

Dilution
A Novel Approach In Preserving Privacy

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von:

Christian Gorecki
aus Dortmund

Mannheim, 2013

Dekan: Professor Dr. Heinz Jürgen Müller,
Universität Mannheim
Referent: Professor Dr. Felix Christoph Freiling,
Friedrich-Alexander-Universität Erlangen-Nürnberg
Korreferent: Professor Dr. Thorsten Strufe,
Technische Universität Darmstadt

Tag der mündlichen Prüfung: 30. April 2013

ABSTRACT

Protection of privacy is a very personal matter and therefore a sensitive issue. Often protection or prevention of exchange of information is crucial to preserve privacy. With information technology on the rise, exchange of information got boosted and preserving privacy turned to a very challenging issue. Commonly, privacy is often understood as non-disclosure of information. Modern media, particularly the Internet, and development of Web 2.0 within the Internet, pose new challenges to the intention of not disclosing certain information for quite a while already. Still, we observe that state of the art is classifying personal information into very few categories - often only two: *visible to friends only* and *visible to everybody*. This does not mirror physical life and the behavior in communication between two individuals.

In this work we move away from privacy by secrecy towards privacy by dilution. Adding enough data to some information under consideration will make it hard to distinguish and hence reveal the information being protected. Dilution is applicable for any kind of data: while in case of plain text additional text can be inserted into the existing text, dilution of pictures and videos is adding additional files of the same type. Furthermore, we enable presentation of different partial identities to different requesters, e.g., a visitor of a web site. Beside a survey that allowed us to derive a basic model here, we elaborated our concepts into two directions. These can be distinguished by their transparency, i.e., the required user-interaction. We introduce *active* and *passive dilution* respectively. Means to efficiently monitor an online reputation, as well as assessments and use case studies regarding robustness, have been conducted. Conclusively, we will see that the dilution methodology is a promising approach pointing to a novel direction in privacy enhancing technologies.

All tools and frameworks presented in this work and contributed by us have been implemented as fully working proof-of-concepts.

ZUSAMMENFASSUNG

Der Schutz der Privatsphäre ist ein sehr persönliches Anliegen und genau darum von großer Bedeutung. Dabei spielt oft das Verhindern oder Schützen von Informationsaustausch eine entscheidende Rolle. Mit dem Einzug der Informationstechnologien nahm die Geschwindigkeit im Informationsaustausch rapide zu und der Schutz der Privatsphäre wurde zu einer großen Herausforderung. Traditionell und im allgemeinen Sprachgebrauch wird Privatsphäre häufig mit der Geheimhaltung von bestimmten Informationen gleichgesetzt. Moderne Medien, hier ganz besonders das Internet und dessen Web 2.0 Entwicklung, stellen uns vor neue Herausforderungen, wenn es darum geht bestimmte Informationen geheim zu halten. Dennoch beobachten wir nach wie vor, dass aktuelle Ansätze sich damit begnügen, personenbezogene Daten in wenige Kategorien zu unterteilen - oft werden dabei nur zwei unterschieden: *nur für Freunde sichtbar* und *sichtbar für alle*. Dies entspricht nicht der physischen Welt und spiegelt das Kommunikationsverhalten zwischen zwei Personen nur unzureichend wider.

In dieser Arbeit möchten wir uns von der Privatsphäre garantiert durch Geheimhaltung wegbewegen hin zu Privatsphäre durch Verwässerung. Fügt man genügend viele Daten zu einer (Menge von) Information(en) die man schützen möchte hinzu, wird es schwieriger, die zutreffenden von den unzutreffenden Informationen zu unterscheiden. Dieser Ansatz lässt sich für jede Art von Daten umsetzen, auch wenn es dabei Unterschiede in der Art und Weise der Verwässerung gibt: Während bei Texten zusätzliche Worte leicht eingefügt werden können, bietet es sich bei Film-, Bild- und Audio-Dateien an durch das Hinzufügen weiterer Dateien zu verwässern. Weiterhin ermöglichen wir mit den Ergebnissen unserer Arbeit das Erstellen von verschiedenen Teil-Identitäten (einer Person), die in Abhängigkeit von dem der die Informationen anfragt (z.B. der Besucher einer Webseite) dargestellt werden. Neben einer Umfrage die es uns ermöglicht hat ein Grundmodell für unser Vorgehen abzuleiten, haben wir unser Kernkonzept in zwei Richtungen ausgearbeitet. Diese können am Grad ihrer Transparenz, d.h. an der Menge notwendiger Benutzerinteraktion, unterschieden werden. Entsprechend führen wir die

Begriffe *aktive* und *passive Verwässerung* ein. Neben Mitteln zur effizienten Überwachung eines “Online-Rufes” stellen wir auch eine Auswertung der von uns festgestellten Defizite, sowie eine Studie zur Robustheit der von uns präsentierten Lösung vor. Zusammenfassend können wir festhalten, dass die von uns vorgestellte Verwässerungsmethode ein viel versprechender Ansatz ist, der eine neue Richtung im Bereich der Technologien zur Verbesserung der Privatsphäre aufzeigt. Alle Programme die wir vorstellen wurden als voll funktionsfähige “Proof-of-Concepts” umgesetzt.

ACKNOWLEDGEMENT

There is one person working on a certain topic, but there are many who ask questions, make suggestions, share their thoughts, provide input, or even do some implementation. Thank you to everybody who did so. Nevertheless, there are some people who took a role of particular importance.

First of all, I want to thank Professor Dr. Felix C. Freiling, who pointed me to this research area. He has been my mentor for more than 6 years and strongly supported me not only in questions regarding this thesis. Next, thank you to Professor Dr. Thorsten Strufe, who's feedback helped me a lot to present the research I did, in a way that it is both, settled in the research context and understandable to general computer scientists.

For proofreading special thanks to Philipp Trinius, who has read the entire thesis and provided a lot of valuable feedback to me. Further thanks for additional proofreading go to Carsten Willems (Chapter 2 and Chapter 3), Markus Engelberth (Chapter 4, Chapter 5, and Chapter 6), and Hans-Jörg Dilzer (Chapter 1).

Thanks to all my colleagues at the Laboratory for Dependable Distributed Systems who created such a great work environment and to all the students I have supervised during my time at the University of Mannheim. Particularly, I appreciate all the inspiring discussions that often led me to new directions in my research.

Most of all I want to thank my wife Jonna, because she had to go through ups and downs of life with me after work hours, my parents who supported me throughout my entire education, and my son Lion who spent quite some hours playing silently in my home office room, while I have been writing the last pages of this thesis.

Contents

1	INTRODUCTION	1
1.1	Motivation & Outline	1
1.2	Prior Art in Privacy Protection	2
1.2.1	Proxying	2
1.2.2	Dilution	4
1.2.3	Online Reputation Management	6
1.2.4	Platform for Privacy Preferences (P3P)	6
1.2.5	Data Minimization/Avoidance	7
1.3	Contribution	8
1.4	Delimitation	10
1.4.1	Terminology	10
1.4.2	Chaffing and Winnowing	11
1.4.3	Threshold Cryptography	12
1.5	Publications and Supervised Theses	13
1.5.1	Related to Dilution	13
1.5.2	Unrelated Work	16
1.6	Conclusion	18
2	IDENTITY	19
2.1	Introduction	19
2.2	Definition	20
2.3	Online Appearance	23
2.3.1	Active Data Traces	23
2.3.2	Passive Data Traces	24
2.4	Linkage	25
2.5	Threats	27
2.6	Conclusion	29
3	SIGNIFICANCE OF DIGITAL PERSONAL DATA	31
3.1	Introduction	31
3.2	Survey	32

3.2.1	Compilation	32
3.2.2	Accomplishment	35
3.2.3	Results	35
3.2.4	Identifying Clusters of Different User Profiles	47
3.2.5	Conclusion of the Survey	50
3.3	Online Profiling	50
3.3.1	Passive and Active Online Profiling	51
3.3.2	E-Recruitment	51
3.3.3	Robustness and Reliability of Online Profiling	53
3.3.4	Monitoring Online Reputation	55
3.4	Conclusion	61
4	DILUTION	63
4.1	Introduction	63
4.2	Definition	63
4.3	History Review	64
4.4	The Idea	66
4.5	Design Concept and Implementation Pillars	67
4.5.1	Initialization	67
4.5.2	Publication of Diluted Identity	69
4.5.3	Request Analysis	69
4.5.4	Partial Identity Composition	74
4.5.5	Delivery of the Composed Partial Identity	79
4.5.6	Results	79
4.6	Application	81
4.6.1	Online Social Networks	81
4.6.2	Personal Homepages	87
4.7	Functional Evaluation	91
4.7.1	Limitations	94
4.8	Conclusion	94
5	EVALUATION OF DILUTION	97
5.1	Introduction	97
5.2	Attacking Passive Polymorphic Dilution	98
5.2.1	Strategy	98
5.2.2	Results	99
5.3	Attacking Active Polymorphic Dilution	103
5.3.1	Strategy	103
5.3.2	Results	104
5.4	Conclusion	105

CONTENTS

iii

6	CONCLUSION	107
6.1	Contribution	108
6.2	Future Work	109
A	Related Solutions, Products, and Services	119
B	Survey	121
C	Interview Questions	129

List of Figures

3.1	Age distribution among survey participants.	36
3.2	Job distribution among survey participants.	37
3.3	Online time spent privately.	38
3.4	Online time spent business related.	38
3.5	Sensitivity to not disclose personal information.	39
3.6	Estimated significance of online reputation within business. . .	40
3.7	Hits among top-ten search results using Google.	41
3.8	Amount of additional pages found searching for the nickname. .	41
3.9	National frequency of surname within Germany.	42
3.10	Comparing Google hits with regards to surname.	42
3.11	Distribution of online social network usage among participants.	43
3.12	Privacy awareness among participants below a given age. . . .	44
3.13	Privacy awareness with regards to age ranges.	44
3.14	Validity of personal information in OSNs.	45
3.15	Google image search hits.	46
3.16	Impact of surname popularity.	48
3.17	Methodology in imaginary identity creation.	54
3.18	Architecture of the Online Reputation Monitoring Framework.	57
3.19	Work flow of online reputation monitoring Firefox add-on. . .	61
4.1	Partial Identity Composition Decision Tree.	75
4.2	Two dimensional privacy graph.	80
4.3	Interactive dilution in OSNs.	83
4.4	Overview Profile Generator.	88
4.5	Profile Generator user input form.	89
4.6	Snippet of Google's cache content.	93

List of Tables

1.1	Privacy solutions based on proxying.	3
2.1	Results of a survey conducted by Acquisti and Gross.	26
4.1	Search engine query URLs.	71
4.2	User-Agent substrings to identify search engine bots/crawlers.	72
4.3	Identity attributes (<i>ia</i>) and their weights (<i>w</i>).	77
4.4	Thresholds for identity attributes.	78
4.5	Categorization and ranking of different identity attributes.	84
4.6	Discrete separation of different validity levels.	84
4.7	Fake profiles database table.	85
4.8	Extension of <i>fusion_user_groups</i> table by an <i>group validity</i> field.	85
4.9	Validity database table.	86
4.10	Simple weighting for personal homepage profile generator.	90
4.11	Thresholds for partial identity composition.	90
5.1	Imaginary identity attributes used as real partial identity.	98
5.2	Results of brute-force attack with five different strategies.	100
5.3	Top 20 after frequency analysis of brute-forcing results.	104

Chapter 1

INTRODUCTION

1.1 Motivation & Outline

There are many research projects focusing on privacy in terms of secure and privacy aware data storage, e.g., peer-to-peer infrastructure using encryption instead of central servers controlled by a not necessarily trusted authority [17, 2]. Thus, concerns regarding the security of data stored in the back-end is not considered in this work. Instead, we consider representation of personal data in order to provide an improved scalability, which is not only based on the decision whether certain information is supposed to be public or private, but instead allows for a more fine-granular distinction between which information to expose and which to preserve private. This way we enable treatment of *digital privacy* as it is done in physical interactions among different people in daily life. Here as well, we usually do not choose between either revealing all our personal information or keeping it all secret, but instead we share different certain subsets of personal details with different parties, e.g., individuals, friends, organizations, etc. Thus, quality and quantity of the information we share strongly depends on the communication peer, i.e., the entity requesting such information. In order to enable this we introduce *dilution*, which forms the base for our contribution.

The remainder of this work is outlined as follows: After a brief introduction to the domain and relevant prior art (Section 1.2), we outline our contribution in Section 1.3 and delimit it from previous contributions, which we found to be most similar to our approach in Section 1.4. An overview on publications that did not relate to this work and the supervised theses contributing to this work is given in Section 1.5. Basic definitions as we use them in the subsequent chapters are to be found in Chapter 2. In this context we also point out most relevant threats our results are aiming for

to mitigate. Chapter 3 will help us to understand the significance of digital, personal data, by first presenting the results of a survey we have conducted and then discussing different aspects on the impact of such data on our lives. In the following Chapter 4 our main contribution, i.e., the design of a novel privacy preserving methodology, is presented. Here, we extend our set of definitions (Section 4.2), look at related phenomena we observed in the past (Section 4.3), and present a design concept as detailed as required for implementation into applications (Section 4.5). Along two example applications are provided (Section 4.6). These we implemented as fully working proof-of-concepts, which are functionally evaluated at the end of the chapter. Looking from an adversary’s perspective an evaluation of the robustness of our approach is given in Chapter 5. Last we conclude on our contributions and results in Chapter 6.

1.2 Prior Art in Privacy Protection

Nowadays, there are different approaches in protecting personal data within the World Wide Web. Many of these address anonymity or pseudonymity on a network layer level as for instance the Tor Project [83] and the Freenet Project [69]. Here, we rather consider solutions in terms of privacy regarding identity (attributes), but we will also refer to other approaches briefly in the following. While we desist from presenting all the solutions, services, and products we have evaluated the complete list can be found in Appendix A.

1.2.1 Proxying

As we will see in Section 3.3, personal data do not only cover actively published, personal information, e.g., identity attributes. There are loads of data, like IP-address, User-Agent, etc. which may be recorded and analyzed whenever somebody is browsing the Internet.

Therefore, different services have been established in order to hide those meta data and thus prevent linkage to a certain dial up-account, i.e., a certain end-user. These services usually base on the idea of proxying, i.e., forwarding requests (and the corresponding responses), without revealing the entity initiating the origin request. In Table 1.1 we list some of these services along with a reference to the corresponding web sites.

While the common understanding of proxy techniques involves three parties, i.e., the entity requesting, the entity responding, and the entity in between (first forwarding the request and then the response), Tor is a little more complex. The **Tor**-Project [83] is employing *onion-routing* in order to not

Table 1.1: Privacy solutions based on proxying.

Product	URL
Anonymouse.org [4]	http://anonymouse.org
Ixquick [81]	http://ixquick.com
Picidae [87]	http://picidae.net
Scroogle [12]	http://scroogle.org
Proxomitron [54]	http://www.proxomitron.info
Privoxy [20]	http://www.privoxy.org
Tor [83]	https://www.torproject.org

allow linkage of a web site request to the origin, namely the person behind the request. We are not going to explain onion routing here, therefore we refer to the cited reference. Important to understand is that onion routing protects privacy on a network layer, whereby it is not possible to trace back the original IP-address of the requester and hence the (dial-up) connection used for the request. Still submitting data, e.g., name or phone-number, to a web site while using Tor is likely to cause a privacy issue.

The **Scroogle** search service [12] unfortunately is no longer online. It used to offer a proxy for Google search queries to render search profiling by Google [36] impossible. As a result queries of several different users of Scroogle were submitted to Google by Scroogle. Thus, Google was not able to distinguish the different users behind the queries. Still, the terms submitted as search queries are visible to Google. Therefore, this is a session layer privacy measure, which does not protect on a data layer, e.g., where users might search for their own name, address, etc.

In order to bypass content-filtering, e.g., by your Internet service provider or any intermediate proxy, **Picidae** [87] offers a service on their web site where you can enter a URL¹. Once you submit the URL, Picidae will take a screen-shot of the target web site the submitted URL is pointing to and return the resulting image (in PNG² format) to you. Using the HTML³ *map* tag the returned image is overlayed with click-able areas right there where hyper references are, i.e., links, which can be found in the original target page. Clicking on one of these areas will in the same way return an image of the corresponding target web site. This way surfing the web is enabled,

¹Uniform Resource Locator

²Portable Network Graphics

³HyperText Markup Language

even though only images are transmitted along with some additional HTML code. The content of the web site is not transmitted as text at any time. Therefore, common content filtering techniques are rendered to be useless. As in all proxy based approaches, the original requester remains invisible to the target server, which only sees requests coming from Picidae.

Proxomitron [54] and **Privoxy** [20] are traditional proxy solutions with additional privacy protecting features, hooking in on a network layer. **Anonymouse.org** [4] provides an analogue function on the application layer via a web site. All three of these solutions will hide the original requester from the target web site. Still these solutions only forward the requests and responses, so that all content of the web sites requested is transmitted over the line and may become subject to content-filtering solutions in place.

We do not want to leave unmentioned that there are research efforts dedicated on proxy approaches applied to online social networks. Looking at the work by Felt and Evans in 2008 with the title “Privacy Protection for Social Networking Platforms” [32], or by Egele et al. titled “PoX: Protecting Users from Malicious Facebook Applications” [24] creates an interesting understanding on how proxying can be applied particularly tailored to the domain of online social networks.

1.2.2 Dilution

Even though the main objective of this thesis is to present dilution as a new approach in privacy enhancement technologies, there already are some approaches following strategies we attribute to the field of dilution. According to our understanding dilution is something very intuitive and therefore rather naturally to appear. Still, there is a difference in doing something directed by intuition, which appears to work out, compared to understanding the idea of a new concept, providing formal definitions, and developing solutions following a concept reasonably.

We first have a look at three approaches, which are somehow similar to the idea of proxying, but still differing in a way that makes us sorting them to the field of dilution. **BugMeNot** [1], **Spambog** [5] and **Cookie Cooker** [85], are three different approaches where users share account credentials, email addresses, or cookies. While a proxy is forwarding requests for different entities and thus is hiding the origin, i.e., the IP-address of the client, who initiated the request, it does not effect on the account data used by the clients. BugMeNot, Spambog, and Cookie Cooker enable sharing of account data with others, to overcome the linkability between user and account related usage data. This way the service in request can not reliably distinguish different users, as they might use the same account data. By these approaches

user profiling is rendered to be of no sense, in case that a certain account information is used by at least more than one user. For the sake of clarity let us consider the following example: Given a user account A and three different users U_1 , U_2 , and U_3 sharing the account A . Let us assume that user U_1 has assigned account A . Then a service trying to profile the activities of user U_1 by monitoring account A , will actually profile the activities of three different users, namely U_1 , U_2 , and U_3 . Thus, the service provider will end up with an aggregated profile of different users, which appear as one user. In other words the profile of user U_1 gets diluted by activities of user U_2 and U_3 . So we are facing a dilution of accountability.

Another technology of hiding information within other data is **steganography**. The wide diversity in how steganography can be applied, makes it hard to give a precise definition. Looking at the word's Greek origin it translates to something like *hidden writing*. There are many ways of how a message can be hidden within other data or physical items. For example a stitching using different spaces between two stitches may encode a message. A text file or email may encode a hidden message by additional spaces at the end of each row. Here we only give two examples but there are many more. And while this appears to be very similar to what we refer to as dilution, there is a significant difference: In steganography the hidden information is visible only to the sender and receiver, who know a certain secret, i.e., how to extract the hidden information, whereas in dilution all information can be visible at anytime to everybody.

Last we want to mention **k-anonymity** as introduced by Samarati and Sweeney [74, 82]. k-anonymity is a property (or requirement) claiming that any particulars which are disclosed allow linkage to at least k different individuals. Thus, privacy of each individual correlates with the size of k . To build the link between k-anonymity and dilution one could try to explain dilution as a k-anonymity preserving approach with a dynamic k , depending on the given knowledge about the individual in question. However, even though dilution can be implemented in a way fulfilling k-anonymity, we will point out the drawbacks, i.e., risk of unintended disclosure of someone's personal information, later in Section 4.7.1.

Other approaches we classify as dilution try to render network monitoring (as a measure for profiling) ineffective by inducing random network traffic, i.e., cover traffic, via automated Internet browsing or search engine requests. Actually, our research on dilution as a privacy measure, is based on a reasoning about the just mentioned methods, and in particular on **Antiphorm** [44]. The same approach has been implemented in the browser extension **TrackMeNot** [84]. The discussion on dilution will be continued in Chapter 4.

1.2.3 Online Reputation Management

All the ideas on privacy protection, we have presented in the previous sections have one common sense: They focus on protecting privacy of the user browsing the Internet. Latest with the establishment of Web2.0 technologies Internet users do not only consume what is provided by different digital services, but become involved in providing content themselves. Internet users do not only download – they do sharing.

Objectives of sharing are digital data, e.g., music, movies, pictures, games, and other information. With the technical devices on hand users do not only reuse digital data, but generate those themselves. For example current mobile phones do not only provide the means to take digital pictures, or record movies, but often allow sharing of these within the Internet by one click.

Even though almost every Internet platform, which offers the possibility to share personal data, has a disclaimer containing a section on privacy, the technical means deployed are rather driven by legal compliance than by the goal to enforce privacy aware usage of such sites. And all approaches taken so far, have obviously not succeeded yet.

As a result we have loads of content contributed by users. Among these data there are many personal information, not only about the users themselves, but also about other people.

Therefore, keeping control over an aggregated identity is rendered to be very hard or even impossible. Of course, this has also been recognized as a business opportunity by some organizations like for instance Reputation Defender [71]. Hence, you are provided with commercial services which will take care of your reputation. Unfortunately, you have to provide many different personal information, i.e., identity attributes, in order to enable them to find data related to you. Since this is somehow contradictory to what privacy is about, we do not get into details on such services but instead present our own solution later in Section 3.3.4.

1.2.4 Platform for Privacy Preferences (P3P)

As mentioned before, most web services have legal disclaimers or privacy policies announced – often as part of their terms and conditions. Due to their extensiveness these are often not read by users and instead silently accepted. Providing a convenient way of comparing privacy policies of different web services with own preferences is the idea of the “Platform for Privacy Preferences (P3P)” [86]. Here, a user may enter once privacy related preferences. In the future, privacy policies of any web site the user visits will be compared

to the before given privacy preferences as the user has configured them on the P3P. In case of mismatches the user will be informed and may decide whether to visit the web site anyway, or not. If policies match no interaction is required. One drawback in this approach is the need for deployment of corresponding P3P policies by the web site administrators. Without these the automatic comparison will not work.

1.2.5 Data Minimization/Avoidance

The idea behind data minimization or data avoidance is to demand, publish and use as less personal data as possible [77, 60]. Obviously, nobody can disclose real identity information if this information is not available or unknown. Thus, the approach of Data Minimization/Avoidance appears to be rather trivial. We do not discuss how to follow this approach, but instead focus on the problems in this methodology. These can be categorized as follows:

Disclosure of Observed Real Partial Identities. Any real identity attribute, which can be observed by another party without interacting with a given identity can also be disclosed by the same party. To give an example, the only way to not have other parties disclose the color of my hair is hiding it.

Disclosure of Virtual Partial Identities. While disclosure of observed real partial identities can at least theoretical be avoided by not disclosing any real identity attributes, disclosure of virtual partial identities cannot be prevented by any means. Anybody can simply make up imaginary, i.e., virtual, identity attributes, also for another party and publish (disclose) these information. Commonly, this is referred to as *spreading rumors*.

Irrevocability of Disclosing Identity Attributes. Any disclosed identity attribute has to be considered as potentially irrevocable. Once a certain information is published it might have been consumed by another party immediately and thus cannot be protected from further publication/disclosure. Commonly, this is referred to as “information loss”.

All three “techniques” can be arbitrarily combined and hence render the data minimization / avoidance approach ineffective. Particularly, irrevocability of disclosing identity attributes does not even require a third party. As our survey (presented in Section 3.2) will show, younger people tend to be rather careless in disclosing their identity. Awareness often raises when

entering a professional career. Nevertheless, then it is too late and possible effects have to be dealt with.

1.3 Contribution

The present work is not about solving one particular problem, but instead looking to prepare a new ground in the area of privacy enhancing/preserving technologies. The idea under consideration here is preserving confidentiality of information by adding additional data and thus making the data of interest hard to distinguish from similar data being added. The more data is added the harder it is to determine the original data. This is what we refer to as *dilution*. An example: Traditionally, we either decide to disclose certain personal information, or to keep them confidential. This could mean for instance, either exposing a hobby, like football, or not. The dilution approach here could be a statement as follows: My hobby is one out of the following five: handball, football, golf, tennis, swimming. Somebody who knows the right information will still know the right answer here. Others who do not know might have a more or less educated guess depending on the relationship between the one requesting and the one exposing such information. The resulting vision of our contribution is to enable communication in terms of sharing and not sharing personal information in a digital world similar to what we already do in physical life: Sharing different subsets of personal information with different people.

Definitions

In order to prepare our research in this field without being biased by existing approaches, we independently derived our own set of definitions to form a model we can use for further analyses, algorithms, and evaluations. In Chapter 2 we define

- *identity attributes*
- *partial identity*
- *identity*
- *aggregated identity*

Except for the aggregated identity we always distinguish between real and virtual components, which allow us to distinguish between the physical and virtual world, without a prejudice on true or false.

Significance of Personal Data

We conducted a survey on the significance of personal data (as presented in Section 3.2) where we have shown the relationship between demographic/personal particulars and the online presence. This can be exploited in applications distinguishing different user groups, as it is required in our dilution approach to allow recomposition of relevant data to the visitors/requesters. We present our utilization of these findings in Chapter 4.

The resulting impact of online presence particularly when it comes to e-recruitment is crucial. Interviews with interviewers have shown the lack of understanding here, even though the means are already in place: Recruiters, look at online presence of candidates without being aware of the risks, i.e., manipulation, etc. To underline this, we present a case study, where we set up an imaginary identity in form of a curriculum vitae and positioned it in the Internet in a way that all statements presented in it can be confirmed via an online research. While in our scenario (publishing very positive personal data) the impact for an individual is rather positive, the opposite can be the case as well. Reputations of individuals can be entirely destroyed by disclosing wrong information online. As a result online reputation monitoring will be a very prominent topic in the coming years. By now there are a few providers, e.g., Reputation Defender [71], offering such services, but we are missing tools to keep the control in the hands of those who are affected. We present a framework which is implemented in a fully working proof-of-concept and allows for reputation monitoring using different channels, e.g., search engines, community sites, or online social network platforms. As an example we have implemented monitoring of Facebook [30] and a generic approach employing the Google [36] search via a Firefox [61] plugin.

Dilution

Research on the significance of personal data, helps to understand how to monitor data, allows for extended analytics and eventually enabled us to successfully build reputation monitoring frameworks. Addressing the prevention of privacy breaches, we present the concept of dilution. Therefore we introduce the high-level concept and elaborate on it down to implementation level. We cover the entire range from diluting data, publication of diluted data (both to users and search engines), request analysis for each visitor, conclusions on the requester derived from its request, composition of a partial identity as it copies communication in physical world, and the delivery of such a composed partial identity. Proving the feasibility of our approach is done by two example applications, one for a personal homepage

and another within a online social network context. Again both applications are fully working proof-of-concepts.

Robustness

Obviously, the whole methodology is not worth of being presented if it does not meet a certain level of robustness to resist attacks aiming to exploit the system. The assessment as presented in Chapter 5 is what eventually confirmed our expectations from the approach as it is discussed throughout this work.

1.4 Delimitation – Discussing (and Comparison to) Other Related Work

While we already had a brief overview on prior art in Section 1.2, we still did not really compare our approach to existing solutions. Of course, we do not want to skip mentioning research and solutions, which are close to our contribution and want to encourage to review and compare the referred works to further improve dilution.

1.4.1 Terminology

While the terminology used in this work, particularly in Chapter 2, was independently elaborated and throughout our research in this field, we point out the publication “Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management - A Consolidated Proposal for Terminology” by Pfitzmann and Hansen [67], wherein the authors suggest a terminology which is almost identical and at least very similar to our proposed definitions. In the remainder of this section we will discuss both similarities and difference in the definitions.

The definition of *partial identity* is probably the closest match between our definitions and the terminology as suggested by Pfitzmann and Hansen. However, we further distinguish between different kind of partial identities (*real*, *virtual*, and *fake*) as this is essential for our dilution approach. Particularly distinguishing between real and virtual for most of our definitions is essential, since the virtual analogue to a definition of something real is commonly extension by imaginary or fake attributes. This is similar to what Rivest refers to as “chaff” (compare Section 1.4.2).

While Pfitzmann and Hansen define *identity* as particular *partial identities* that allows to “sufficiently identify [an] individual”, we understand identity as the superset of all partial identities, since also virtual or fake partial identities may relate to the individual under consideration and this should be covered to model our approach. Furthermore, we also emphasize that one individual “owning” a certain identity will present different partial identities depending on the situation and context. Such situation or context is defined as *Role* in Pfitzmann’s and Hansen’s publication, where they base this definition on connected actions within a certain social situation. The *aggregated identity* as we define it later is named *complete identity*. For any definitions of *anonymity*, *(un)linkability*, *(un)detectability*, *(un)observability*, and *pseudonymity* we refer to their publication as we do not have a need for formal definitions here. They also look at the *communication* network as a system of *acting entities* and provide corresponding definitions. For our approach this was out of scope, since we do not look at privacy at the network level.

A little bigger gap is between our definitions of *(identity) attributes*, which they define by *characteristics* or *actions*, our *virtual identity*, which they use to describe what we define as *fake partial identity*, and *digital identity*, what we define as *virtual identity*.

Deliberately, we chose *linkage* instead of *linkability* for determination of relation between any form of identities and identity attributes as we also found *linkability* to be preserved for relation between items of interest (to an attacker), e.g., subjects, messages, actions.

1.4.2 Chaffing and Winnowing

The approach which we found to be closest to our dilution concept was presented by Rivest in 1998 [73]. In his article titled “Chaffing and Winnowing: Confidentiality without Encryption” Rivest suggests a technique as described in the following. In order to preserve confidentiality of transmitted data from a sender s to a receiver r , the sender adds additional information, i.e., *chaff*, which is transmitted randomly mixed with the actual payload. In order to allow the receiver to easily distinguish between good data (*wheat*), i.e., data containing the actual information s wants to share with r , and chaff, s authenticates all good data by adding a MAC (message authentication code) using a secret key, pre-shared with r . All other packets, i.e., the chaff, have a random MAC like looking string.

As a result confidentiality of the relevant information (wheat) is protected by being diluted with additional information. Even though this appears to be pretty similar to our approach there are some major differences we would

like to highlight in the following.

Key-Based While the contribution by Rivest was explicitly meant to not employ encryption for preserving privacy, usage of a pre-shared key is central to the idea presented. Both, sender and receiver need to exchange a key in order to allow for authentication of wheat data by the sender and distinguishing it from the chaff by the receiver. Once the key leaks, confidentiality is not preserved any longer. Here, we do not discuss robustness of Chaffing and Winnowing, but the interested reader we refer to an article by Bellare and Boldyreva titled *The Security of Chaffing and Winnowing* [10]. Instead, we assume the algorithm is secure as long as the key remains secret. Thus, a third party can either extract no information at all (if not in knowledge of the secret key) or read the entire information (if the key is known). The dilution approach presented in this work is meant to not employ any keys, but form a decision on how many and which relevant data to present, based on (meta) information extracted during the interaction between a requester and a web site where a certain individual has published its particulars. Note, that the web page might also be a profile page within a online social network.

Legal Aspects Chaffing and Winnowing has been designed with the intention to provide an alternative confidentiality measure, i.e., to keep communication data secret, in order to overcome legal limitations as for instance export regulation on encryption techniques in the USA [78]. Rivest aimed to raise the barrier for intercepting confidential communication by basing his approach on authentication rather than on encryption.

1.4.3 Threshold Cryptography

Another domain, which comes to mind is *threshold cryptography* [19]. The idea is to allow decryption of a cipher only if a certain amount (threshold) of secret keys is available. This way the power of decryption can be spread over or shared by different entities, e.g., individuals. For the sake of simplicity we consider the following example: Assume a message is encrypted using ten different keys. Each of the keys is known by exactly one person and there is no person knowing two different keys. Then decryption of the cipher, i.e., the encrypted message, might for instance require eight out of the ten entities to provide their key. Particularly, it does not matter which eight of the ten keys are provided. This way, having for instance three entities that do not provide their keys, the cipher cannot be decrypted and the original message cannot be restored. While there is a similarity to our thresholds related to

the ranking algorithm we present in Section 4.5.4, our algorithm allows for determining the amount of information to be shared/disclosed, where threshold cryptography will allow for decryption of all encrypted information once a threshold in the amount of necessary decryption keys is reached. Nevertheless, this is only a similar aspect in two different approaches. We use dilution where threshold cryptography uses encryption. We define thresholds on input derived from (meta) information during information exchange between requester and server in order to derive the amount of exposed real information, where threshold cryptography defines a threshold on input in form of keys, necessary to encrypt (and expose) all information.

In this section we discussed and compared terminology, technology and research which can be somehow related to dilution with our contribution. While there are many more privacy approaches and methodologies we tried to focus on those that appear to be most relevant in terms of dilution. Even though the contributions we refer to in this chapter have some aspects in common with our approach, there is still a significant gap. This gap will be obvious, after we have presented our concept in Chapter 4.

1.5 Publications and Supervised Theses

In this section we list publications and supervised theses. Here we distinguish between the works related to this thesis and unrelated contributions. The first mentioned we attribute to the corresponding sections of this work whenever possible. It is in the nature of research that not all projects lead straight to the right directions. Nevertheless, all projects helped to understand and learn about this new area of privacy preserving technology and thus contributed to the overall outcome.

1.5.1 Related to Dilution

Diploma/Master Theses

Verwässerung von Persönlichkeitsprofilen im Internet (Sinem Kuz, 2009) [52] The title of this work translates to *Dilution of User-Profiles in the Internet*. This work founded the base for our design concept as presented in Chapter 4, particularly on passive dilution as outlined in Section 4.5.3 and the corresponding implementation as described in Section 4.6.2.

Clusterbasierte Analyse zur Internetpräsenz von Personen (Pascal Göbel, 2010) [35] In English: *Cluster-based Analysis of People's Internet Presence*. Within this work the survey as presented in Section 3.2 has been conducted. This work was based on a bachelor thesis by Boris Margara [58] (see below). The aim was to improve the previous survey in order to allow for an analysis of relationship between demographic information and the online presence of an individual.

Reputation Monitoring - Entwicklung eines halbautomatischen Systems zur Überwachung der eigenen Internetpräsenz (Alexander Juhn, 2011) [48] In English: *Reputation Monitoring - Development of a Semi-Automated System to Monitor the own Internet-Presence*. The results of this work contributed particularly to Section 3.3.4, where the resulting online reputation monitoring framework is presented. The main purpose here was the development of a solution which can be run independently by any individual in order to not rely on privacy policies of providers offering this as a service, e.g. Reputation Defender [71]. The modular design allows to integrate further online social network sensors, etc. in a plugin fashion. In the proof-of-concept implementation a plugin for Facebook [30] is given.

Bachelor Theses

Identifizierbarkeit im Internet: Zur Signifikanz personenbezogener Daten (Boris Margara, 2009) [58] In English: *Identifiability in the Internet: About Significance of personal data*. This work was the first attempt to better understand the impact of different identity attributes, when being found/searched on the Internet. Due to the early stage of our research the learnings from this effort formed the main contribution since they supported the survey as conducted by Pascal Göbel [35].

Verwässerung des persönlichen Profils in Sozialen Netzwerken (T. Dang Duc, 2010) [23] In English: *Dilution of User-Profiles within [Online] Social Networks*. This work contributed to the concept of active request analysis as presented in Section 4.5.3 and resulted in a proof of concept implementation within the context of online social networks as shown in Section 4.6.1.

Online Reputation Inspection: Continuous Monitoring the Online Presence of People (Katharina Reich, 2011) [70] Within this work a Firefox add-on has been developed to monitor the own online presence

via periodic Google queries. The proof-of-concept is working in a stand-alone fashion but is prepared to interface with the framework as presented in Section 3.3.4.

Angriffe auf dynamische Profilgeneratoren (Johannes Grohmüller, 2011) [40] In order to prove the robustness of our concept this work titled *Attacking Dynamic Profile-Generators* was looking from an attackers perspective for bypassing the privacy protection offered by our concept. The result as presented in Chapter 5 shows that our method is not only functional but even robust against attacks.

Student Research Projects

Datenspeicherung in Sozialen Netzwerken (Christoph Bales) [8] In English: *Data Storage in [Online] Social Networks*. This work was addressing state-of-the-art in online social networks: What privacy measures are in place, what are the terms of conditions, etc.

E-Recruitment (Nina Sophie Stadler, 2010) [80] Talking to different human resource departments this study helped us to understand the role of online presence in general and its particular role in recruitment.

Einblicke von innerhalb und außerhalb in Soziale Netzwerke (Konrad Nuhn, 2010) [64] In English: *Looking at online social networks from the perspective of an outsider or member*. This work pointed out that even though some information is meant for members only, information leakage is still happening. During our research, we saw at least two publications [11, 63] confirming this and also raising awareness, so that many online social networks improved their privacy settings in terms of enforcement of policies.

Fiktive Identitäten in Sozialen Netzwerken: Chancen und Risiken im Bereich Recruiting (Daniel Köhler, 2010) [51] In English: *Imaginary Identities within Social Networks: Chances and Risks Particularly in Recruiting*. This work founded the base for Section 3.3.3. By creating a resume and publishing it along with relevant information in the Internet in a way that somebody validating the resume via an online survey will be confirmed in believing the story, we demonstrated the risk of trusting digital identity information. This is particularly due to the fact that digital

information are much easier to spread than without a corresponding ICT⁴ infrastructure.

Alternative Ansätze zum Schutz der Privatsphäre (Christoph Klau, 2010) [50] in English: *Alternative Approaches in Privacy Protection Techniques*. This work supports Section 1.2.

1.5.2 Unrelated Work

Beside our research in the area of privacy we engaged particularly in research on malicious software analysis and digital forensics. Corresponding publications and supervised theses, which have not contributed to the present work are listed in the following. As these works do not relate to the present work we list them without further comments.

Publications

Measuring and Detecting Fast-Flux Service Networks [45] (Thorsten Holz, Christian Gorecki, Konrad Rieck, Felix Freiling) This paper was presented at NDSS 2008⁵ and proposed the first detection technique to efficiently identify fast-flux domains.

Das Internet-Malware Analyse-System (InMAS) [27] (Markus Engelberth, Felix Freiling, Jan Göbel, Christian Gorecki, Thorsten Holz, Ralf Hund, Philipp Trinius, Carsten Willems) This article published in the journal *Datenschutz und Datensicherheit* presents the Internet malware analysis system as developed and integrated at the University of Mannheim. The system includes malware capture, analysis, clustering, and visualization.

The InMAS Approach [25] (Markus Engelberth, Felix Freiling, Jan Göbel, Christian Gorecki, Thorsten Holz, Ralf Hund, Philipp Trinius, Carsten Willems) This paper was published at the ENWI 2010⁶ and focuses on capturing and analysis of malware as implemented in InMAS.

⁴Information communication technology.

⁵<http://www.internetsociety.org/events/ndss-symposium>

⁶<http://www.enisa.europa.eu/events/enisa-events/ENWI2010>

Frühe Warnung durch Beobachtung und Verfolgung von bösartiger Software im Deutschen Internet: Das Internet Malware-Analyse System (InMAS) [26] (Markus Engelberth, Felix Freiling, Jan Göbel, Christian Gorecki, Thorsten Holz, Philipp Trinius, Carsten Willems) This work was presented at the BSI congress Deutscher IT-Sicherheitskongress ⁷ and focuses on malware capturing and analysis by presenting various measurements and statistics accordingly.

Mail-Shake [28] (Markus Engelberth, Jan Göbel, Christian Gorecki, Philipp Trinius) This paper, presented at DEXA 2009⁸, suggests a new approach in mitigating unsolicited emails without altering the email protocols in use. Instead a handshake is proposed, which can be conducted by the user and thus build on top of the existing email protocols.

TrumanBox - Transparente Emulation von Internetdiensten [38] (Christian Gorecki, Felix Freiling, Marc Kühner, Thorsten Holz) This paper was presented at SSS 2011⁹ and describes a transparent system emulating different Internet services in order to allow dynamic malware analysis without Internet connectivity and still providing the necessary interaction to monitor malicious behavior.

Theses

- Analyzing Fast-Flux Service Networks (Patrick Scharrenberg, 2008) [75]
- MailShake Mailclient Plugin (Martin Gräßlin, 2010) [39]
- Forensische Datenanalysen mit Data Mining (Michael Riecker, 2009) [72]
- Modellierung und Analyse von Fraud in elektronischen Geschäftsprozessen (Alexander Pfister, 2009) [66]
- Analyse des Vorgehens und Verhaltens von Spammern und Harvestern bei dem Auffinden von E-Mail Adressen im Internet (David Passarelli, 2010) [65]

⁷<https://www.bsi.bund.de/SharedDocs/Termine/DE/2013/13DeutscherITSicherheitskongress.html>

⁸<http://www.dexa.org>

⁹<http://www.jaist.ac.jp/ss2011/>

1.6 Conclusion

After motivating our research we discussed prior art technologies including proxying, online reputation management, P3P (privacy preserving platform), and data minimization. We also had a look at existing technologies we categorize as dilution. Still these are not understood as such. Thus, building novel privacy technologies that employ dilution requires a proper understanding of this research area and example applications showing the effectiveness. Both is provided in the remainder of the present work. Furthermore, we summarized our contribution and delimited it from related works. A list of publications and supervised theses is provided and attributed to the corresponding sections in the remainder of this work wherever feasible.

Chapter 2

IDENTITY

2.1 Introduction

In this chapter we discuss the term *identity* and its particular meaning within the context of the World Wide Web. Therefore, we consider both the traditional understanding of identity within the physical world and its counterpart in the digital world. In order to understand the meaning of identity in the Internet, we discuss differences and similarities of both forms of identity. As our main objective is preservation of privacy we will have a look on risks which may arise within a digital context.

Thinking of real world scenarios, we can observe varying behavior of a certain person in different situations. Since behavior can be seen as a mirror of personality (compare Manoharan’s book “Education And Personality Development” [57]) and hence of identity, we may understand these appearances as reflections of *partial identities*. These partial identities in term consist of personal attributes (*identity attributes*), which in conjunction form the (*aggregated*) *identity* of a certain individual. Partial identities may differ from the aggregated identity in a way that all personal attributes of the aggregated identity may vary in their characteristic for each partial identity. As we can see, there are many different terms involved when speaking about identity: For the sake of clarity we will provide formal definitions in Section 2.2.

According to the right to privacy each individual may decide who knows about which of its partial identities. Facing development of the World Wide Web over the last few years, e.g., Web 2.0, the challenge of defending this certain right has been significantly increased. Personal information is spread all over the Web. Given a few information about a certain individual, and using the tools at hand it is easy to gather many personal properties commonly reflecting more than only one partial identity. By using correlation

these properties can be combined to a detailed view on a certain identity [63]. Thus, we end up with a bundle of information enabling us for momentous decisions or actions. The resulting information might be used as a decision base for instance in business relationships, or even being abused in a criminal context.

In the following section, we prepare the stage for in-depth understanding of already known problems regarding privacy, corresponding counter measures, downsides of these counter measures, and our new approach in privacy preserving technologies.

2.2 Definition

In order to have a common understanding on what identity and particularly virtual identity consists of, we will next present some formal definitions. Thus, we can precisely argue appropriately in the remainder of this thesis. Following a bottom-up approach, we first have a look at what an identity consists of and therefore define what an identity attribute is. In the following we only consider individuals and not things because things do not matter in our context. Anyway, in all definitions the term *individual* may be replaced with the term *thing* in order to obtain analog definitions for material items.

Definition 2.1 *An **identity attribute** is an attribute characterizing a certain individual.*

As an example we take John Doe, who has blond hair, green eyes, weighs 80kg, and has reading as a hobby. The identity attributes here are “John”, “Doe”, “blond”, “green eyes”, “80kg”, and “reading”. Depending on the situation, different identity attributes may turn visible. For example in a conversation we may decide which identity attributes to reveal. The resulting phenomenon we call a *partial identity*.

Definition 2.2 *A **partial identity** \mathcal{P} is a finite set of identity attributes characterizing the very same individual:*

$$\mathcal{P} = \{i_1, i_2, \dots, i_n\}, \text{ with } n \in \mathbb{N}$$

A partial identity can be thought of as a view on a certain part of an identity. Reviewing our previous example one partial identity of John is: “John” and “blond”. Accordingly, the corresponding identity is a superset of all related partial identities.

Definition 2.3 An **identity** \mathcal{I} is the superset of all partial identities \mathcal{P}_k , with $k \in \mathbb{N}$, characterizing the very same individual. Thus, we also write $\mathcal{P}(\mathcal{I})$ to refer to an partial identity of \mathcal{I} . Formally:

$$\mathcal{I} \text{ is identity iff } \forall \mathcal{P}_k(\mathcal{I}) : \mathcal{P}_k(\mathcal{I}) \subseteq \mathcal{I}$$

Already in the physical world a certain individual might be connected to different identities. For instance an actor has its own personal identity, but while performing a certain role in an act the actor usually performs with a different identity, i.e., the identity of a certain character. Thus, we have to distinguish between reality and virtuality.

Definition 2.4 A **real identity attribute** is an identity attribute truly characterizing a physical individual in physical life.

Definition 2.5 A **virtual identity attribute** is any identity attribute, which is not a real identity attribute.

In general, a partial identity may consist of both real and virtual identity attributes. If all identity attributes of a given partial identity are real identity attributes then we refer to the given partial identity as *real partial identity*. Otherwise we call it *virtual partial identity*. Accordingly, we may distinguish between *real* and *virtual (partial) identity* as follows:

Definition 2.6 A **real partial identity** is a partial identity exclusively consisting of real identity attributes.

Definition 2.7 A **virtual partial identity** is any partial identity which is not a real partial identity.

Therefore, pretending or claiming to be someone you are not means creating a new virtual identity derived from your real identity. It is important to note that in our understanding of virtual identities those are not restricted to the digital world, but may also occur in daily, physical life.

Definition 2.8 A **real identity** is an identity exclusively consisting of real identity attributes.

Definition 2.9 A **virtual identity** is any identity which is not a real identity.

While in many scenarios a virtual identity might match a physical identity, e.g., in online social networks, they might also differ. An online role play character for instance might have nothing in common with the physical individual actually controlling the character.

Anyhow, the role play character might have certain attributes in common with the physical individual controlling the character. In the latter case we are facing a mixture where an identity attribute might be a virtual identity attribute (for the role play character) and a real identity attribute (for the individual behind the role play character) in the same time.

Furthermore, the same identity attribute might belong to different individuals. But for two different individuals with real identities \mathcal{I}_1 and \mathcal{I}_2 it always holds

$$\mathcal{I}_1 \neq \mathcal{I}_2,$$

which is equivalent to

$$\exists i_x : (i_x \in \mathcal{I}_1 \wedge i_x \notin \mathcal{I}_2) \vee (i_x \notin \mathcal{I}_1 \wedge i_x \in \mathcal{I}_2).$$

In other words, there is always a real identity attribute by which two different physical individuals can be told apart. This does not necessarily hold for virtual identities.

Even though our definitions already cover trustworthiness – *real identity attributes* always refer to true facts – we stress that a virtual identity might be completely imaginary. This is where we turn to a new aspect of understanding identities. A certain individual might not only have different identities it binds to but also initiate or create independent virtual identities. In the first place those might appear to be irrelevant for understanding the real identity of the individual causing them. Still, those virtual identities might mirror certain aspect of the real identity behind the scene and often do disclose otherwise hidden real identity attributes. Therefore, it makes sense to consider both, the real identity as well as all virtual identities related to a certain physical individual, in order to collect information, i.e., real identity attributes, regarding this certain individual. By virtual identities we refer to all virtual identities and not only those, which obviously mirror certain real identity attributes.

For this purpose we define an *aggregated identity*.

Definition 2.10 *An **aggregated identity** is the union of all (real and virtual) identities related to or created by a certain individual. Let \mathcal{I} be the identity of a certain individual then we write $\mathcal{A}(\mathcal{I})$ to refer to the aggregated identity regarding \mathcal{I} .*

In the physical world, virtual identities, apart from characters in an act, are rather uncommon. In contrast the digital world renders creation of new virtual identities to become much easier. The increasing amount of people joining online social networks in order to create virtual mirrors of their physical identities is tremendous. And this phenomenon is not restricted to online social networks. All different kind of Internet platforms denote a rapidly rising number of users.

For the sake of simplicity in describing different scenarios within the remainder of this work we also define a particular virtual (partial) identity, i.e., the fake partial identity.

Definition 2.11 *A **fake partial identity** is a virtual partial identity with no real identity attributes at all.*

We will extend this set of definitions in Section 4.2, where we discuss dilution. In the meantime the given definitions will be sufficient to formally describe our research results.

2.3 Online Appearance

Even though there are various impacts one cannot control in order to preserve a positive online reputation or keep personal data undisclosed, the majority of personal information is provided by each Internet user itself. Some of this information is shared intentionally to increase the own visibility or build up some reputation. Other information is published without consideration of the impact this information might have. Both kinds of data are actively published by the user and hence are referred to as *active data traces*. Apart from these there is another category of information provided by the user itself: *passive data traces*. Still there is no third party involved, but the sites a user is interacting with, i.e., browsing. In this section we briefly describe the difference between active and passive data traces, since this will form the base for our implementations presented in Section 4.5.

2.3.1 Active Data Traces

15 years ago, actively publishing personal information commonly required having own web space available. Therefore, it was necessary to have a contract with some web space provider. Even though there were web spaces for free (financed by commercial advertisement) quite from the beginning, serious web presence, i.e., without advertisements, involved renting web space

for a certain fee. Shared platforms where different individuals publish information have been limited mainly to dedicated communities. In this environment keeping track of the published information was rather easy - at least compared to the current situation. With the introduction of an idea which is commonly referred to as Web 2.0 the scene has changed tremendously. While the (modern) Internet has started as a place where a minority mainly commercial entities provided information meant to be consumed by everybody having access to the Internet, it turned to a participating community, in which almost every consumer becomes a contributor. Online social networks pose the prime example here. All these contributions, e.g., blog posts, forum entries, comments on other people's posts, profile pages, picture galleries, we refer to as **active data traces**. As a result control over active data traces became very challenging if not impossible. Many users increase the amount of active data traces at least on a daily bases for instance by publishing status updates within social online networks: *at home, at work, on vacation, work out, watching TV*.

Thus, tracing activities of a certain individual becomes easier the more of such information is available. Additionally, also other users might contribute to the active data traces of a certain individual. They might post information like: *at home with ... , going to movies with ... , etc.* Obviously, this renders preservation of privacy to become even harder and correlation over all these active data traces regarding a certain individual enables a transparency which is commonly not intentional.

2.3.2 Passive Data Traces

Although participating in the information exchange within the Internet in form of contributing personal information of any kind is easier than ever before, there are also users being more privacy aware. By mainly consuming information available on the web they feel less exposed. In the same time such kind of users wonder why for example an online shop highlights exactly those products the user is looking for, even though the user did not (actively) provide these information and also did not login with credentials that might allow linkability. The previous example describes the effect of personalized web or in particular of personalized advertisements. Evaluating different meta information or exploiting techniques Internet users are often not aware of, it is possible to learn a lot about the visitor of a web site, without asking the visitor to provide such information. Assuming that no additional privacy measures are employed, a web site can extract information like, the preferred language of the visitor, which web browser is used to display the current page, which web page has been visited last, in case the previously visited

web site was a search engine, the search terms can commonly be extracted, when the visitor has been on this web site last.

These are just a very few examples of information available from the “perspective of a web site”. All of these we refer to as **passive data traces**. As a result, personal information is not only shared when a user is actively publishing such data, but also while *silently* browsing the web [79].

2.4 Linkage

As we have discussed in Section 2.2 different partial identities are revealed depending on the context, which involves situation and people. In physical life this decision often corresponds to some locational distance between places where different partial identities are disclosed. Therefore, locational distance can be seen as a natural separator regarding partial identities.

The resulting borders between partial identities provide the actual privacy protection by rendering linkage between different partial identities of a corresponding identity to be a hard problem or even impossible.

In a digital context these borders loose their significance. Still people consider themselves to be in a certain scenario and thus decide about which partial identity to reveal respectively, while not being aware of the context change. Each partial identity placed or published within the digital context of the Internet is accessible by all participants or individuals by default, leaving password protection and other access limiting measures being unconsidered. Thus, the locational distance loses its meaning or rather turns to be of no significance at all. As a result all partial identities may be gathered and correlated in order to determine partial identities belonging to the very same aggregated identity. This way it is possible to enumerate characteristics of the corresponding real identity to a certain extent. In the worst case it might even lead to a full disclosure of all personal properties belonging to the real identity and hence the loss of privacy.

This shortcoming is favored by novel technologies like Web 2.0 and the lack of its users’ awareness caused by experiences within a physical context. While the idea of Web 2.0 causes concerns regarding users’ data in general, there are subcategories focusing particularly on personal attributes related to their identity, namely online social networks (OSNs). We will discuss the application of our approach within the scope of OSNs in detail in Section 4.6.1. In terms of linkage it is sufficient to think of OSNs as platforms enabling their users to publish a partial identity. Given the rapid rise in the amount of OSNs during the recent past, there are social networking platforms covering almost every topic. Since most of the social networks focus on a certain

objective, e.g., education, hobbies, sexual disposition, etc., people tend to not only participate in different social networks but also, reveal different partial identities regarding the topic of a given social network.

Understanding a certain social network identity as a set of identity attributes, overlapping of subsets of identity attributes might lead to linkage of different partial identities by a third party and thus disclosure of the aggregated identity. Feasibility has been proven by Balduzzi et al. [7]. Therefore, particularly online social networks pose a threat for privacy.

In 2006 Allesandro Acquisti and Ralph Gross conducted a survey [3] on which information is published on Facebook. Table 2.1 contains some of their results.

Table 2.1: Results of a survey conducted by Acquisti and Gross [3].

Information	provide	not provide	not accurate
Birthday	12%	84%	3%
Personal address	73%	24%	3%
Cell phone	59%	39%	2%
Home phone	89%	10%	0%
Political views	42%	53%	6%
Sexual orientation	38%	59%	3%

Particularly noteworthy is the fact that almost half of the participants do not have any concerns about providing their political attitude and even 73% publish their private address.

Besides online social networks like Facebook, there are plenty of other options where users might publish personal information. Bulletin boards, blogs, online market places, or personal homepages are commonly used for self-expression, exchanging information, or for using services like online shopping or information services. In general, personal data are provided without any concerns by users having certain addressees in mind: Who else might visit and read a bulletin board on photography, than those who are interested in the topic themselves. Joshua Fogel and Elham Nehmad conducted a study [33] showing that 73.6% of the participants have worldwide accessible profile data within online social networks, since they do not expect access by foreigners or people they do not want to address anyway. The possibility of automated data gathering techniques, e.g., by using crawlers¹, and correlation of different information using well-engineered algorithms based on mathematical concepts, is something an average user of the Internet is not

¹Crawlers are programs (software) browsing the Internet in an automated fashion.

aware of. This is not the only lack of information Internet users suffer regarding their privacy. A privacy aware user might not want to publish any personal information, but just use the world wide web as an information resource and pool of service providers, e.g., online shops. While trying to publish as little information as possible actively, there are loads of (meta) information being collected passively, i.e., without the users' awareness. IP addresses, visited web sites, search queries, duration of visit, time, the origin web site or search engine the user has been linked to the current web site, geolocation, operating system, browser version, etc. are just a few examples of information being collected in the background. Particularly in online social networks Narayanan et al. [63] and Chew et al. [16] have shown how privacy can be subverted by using selected meta information and mathematics. Availability of cheap storage media is the reason for the long-life cycle of all these data [42, 76].

There are already different approaches trying to prevent or lower the amount of valuable information collected passively, some of which we referred to in Section 1. An approach to prevent correlation of information published actively we will present in this work, particularly in Section 4.4 and Section 4.5. Still we remain with the problem of raising awareness of the average user, who is not familiar with the technique behind the web sites visible in the Internet.

Here, the most efficient approach might be illustration of online presence and its effect, instead of technical explanations. A good example for demonstrating online visibility is given by *Yasni* [93].

As a result, informational self-determination is hard to preserve. Even if an individual would be capable of monitoring personal data published about itself, there is no way of controlling further data processing as for example correlation.

2.5 Threats

In the previous sections we have discussed passive and active data traces and different forms of identities. Additionally, we had a look at linkage techniques, which are particularly useful in finding different partial identities referring to the same individual. As a result we face a threat enabling us to reveal more comprehensive partial identities as intended by the publisher and owner of the corresponding identity. Next, we will turn to general threats impacting our (digital) privacy to form a better understanding on what we protect against. We focus on those threats which are of particular interest within the scope of this thesis.

Harming Privacy and Digital Reputation

Harming privacy and harming digital reputation generally are two different objectives. Particularly, when understanding privacy as *the right of an individual to decide which identity attributes to disclose to whom*, privacy by definition can only be harmed if identity attributes are disclosed without the corresponding individual agreeing on this. Accordingly, publication of fake identity attributes cannot be considered as harming privacy. Thus, harm of privacy can be quantitatively measured, in terms of number of identity attributes (not fake) being published. Harming digital reputation can be done twofold, either by disclosing real identity attributes or virtual identity attributes. In both cases harm on digital reputation can be measured by both quality and quantity of the corresponding identity attributes: An identity attribute shedding bad light on the corresponding individual can be considered as harm on its digital reputation. Depending on the particular information the harm can be of different quality. Additionally, the amount of identity attributes has a direct impact on the reputation. Here, we again have a quantitative measure. Hence, the digital reputation of an individual can be understood as the overall sum of qualities of all different identity attributes being disclosed. The type of identity attribute, whether it is real or virtual, does not really make a difference here. This only effects if identity attributes are verified, i.e., tested whether they are real identity attributes. Due to the ease of publishing virtual identity attributes, digital reputation can be seen at high risk.

The impact of digital reputation has been investigated in 2007 by the “Bundesverband für deutsche Unternehmensberater” (engl. Federal Association of German Corporate Consultant) [14]. As a result 30% of 270 participating human resources consultants have claimed to include an online survey into their decision on candidates. In 57% of this surveys the result lead to rejecting the candidate.

Data Trading

Another discipline being raised to a new level digital age is data trading. Identity data have been known to be valuable from an economic perspective for a long time. Whenever sharing such information there is a risk on having this information being traded to agencies in order to raise added value for different marketing campaigns. While the non-digital form of this information takes much more effort to be processed, the bar has been lowered for its digital counter parts. A part from asking people to agree on sharing this information, identity data can be harvested from the Internet in an auto-

mated fashion, without the affected individuals noticing it. Hereby, control over identity attributes of a certain individual is rendered impossible, and hence privacy cannot be preserved.

Identity Theft

Many online services such as digital market places, online shops, or trading platforms require the user to provide identity information in order to enable accountability. Given that many identities are disclosed online, abuse of such information is the consequence. Collecting required information related to a certain individual enables anybody to impersonate this individual within the digital world. This kind of crime, known as identity theft, enables interacting and doing business in behalf of someone else without his or her affirmation. Beside the loss of accountability this sort of crime can have a serious impact on the reputation of the identity being impersonated. In most cases using a stolen identity for any kind of business or interaction has a negative impact on the identity in use. Otherwise, there would be no sense in taking the risk of getting caught doing business in behalf of somebody else if the same business could have been done with legitimate, i.e., own identity data, without concerns.

2.6 Conclusion

In this chapter we have discussed what an identity is, how it is reflected within the Internet, and the upcoming threats related to identity and privacy of an individual. The definitions presented will be used consistently throughout the remainder of this work. Moreover, the way definitions were developed will support a new paradigm of understanding preservation of privacy by employing dilution. In particular, these definitions help to transfer privacy strategies in daily interaction between different people (in a peer-to-peer fashion, from the physical world) into the digital world.

Chapter 3

SIGNIFICANCE OF DIGITAL PERSONAL DATA

3.1 Introduction

As presented in the previous chapter an online identity in general is formed by different (real or virtual) personal identity attributes published in the world wide web. Significance of a certain personal information is depending on different factors:

- **Type** of personal information, e.g., name, email, phone,
- **(in)validity** of the information,
- **place** of publication, e.g., web site, blog, OSN,
- **privacy measures** in effect, e.g., limited visibility.

Even though this is not a complete list of all influences it reflects major features as considered by search engines and human sense when gathering personal information.

In this chapter we present a survey, which we conducted in order to learn about the relationship between demographic features and online presence/visibility, followed by a discussion on online profiling, where we consider technical means, applications particularly in e-recruitment, robustness, and last present our framework for monitoring online reputation without relying on any third party offerings.

3.2 Survey

In this section we describe the survey's compilation, conduction, and evaluation. Before we get into discussing how the survey has been compiled, we want to mention that this survey is not a controlled study, i.e., meant to measure a certain parameter, among a well defined group, after applying scientific sampling. Instead the idea was to not drive the survey into a specific direction, but to keep it as open and undirected as possible. Our intention here is to eventually have two parts of the survey, i.e., one on demographic questions and another on online presence, where we can cluster the participants according their answers. The participants are once clustered with regards to their demographic features (as deducted from the participants answers) and once with regards to their online presence as evaluated throughout the course of the second part of the survey. Eventually we compare the two resulting cluster sets and map between those. Due to the fact that we did not apply proper sampling among the participants we are aware that this is likely to bias our survey's result. Nevertheless, we do not claim to a generally valid set of clusters any individual can be classified by, but we want to prove that it is possible to derive enough information to cluster online users and thus conclude on their online presence and how to measure this best. Last we want to note that a test survey has been conducted where 260 individuals participated. This helped us a lot to tailor the questionnaire to enable best application of clustering and measures to meet normalization. We desist from presenting the results of the test survey, and instead only present the eventual survey in compilation, conduction and evaluation.

3.2.1 Compilation

As we are interested to learn more about the relation between personal information, behavior, and the resulting online visibility, we decided to cover the following five topics within our survey:

1. General personal information,
2. exposure of personal data in the Internet,
3. online search experiment,
4. social networks, and
5. personal pictures on the Internet.

Each section contains questions asking for further details. Certainly, the answers depend on the estimation of the participants. The online search experiment guides the participants to perform different online searches for certain particulars, so we can validate estimations of the candidates and relate the resulting online visibility to prior behavior and particulars. The complete questionnaire may be found in Appendix B. However, we present the different sections of the questionnaire in the following.

General Personal Information

The first part of the questionnaire is covering personal information. During evaluation the corresponding answers help us in clustering participants into different user categories regarding age, profession, gender, etc. The age is distinguished by certain ranges as follows:

- below 14
- 14 - 17
- 18 - 21
- 22 - 29
- 30 - 49
- above 49

The higher precision within the range between 14 - 29 helps to distinguish between different education stages, particularly in terms of social networks. There, we are faced with social networks with different addressees, e.g., pupils, students, employees.

Exposure of Personal Data in the Internet

In this section the disposition for exposing personal data in the Internet is tested. Beside elaborating which of the identity attributes¹ people are willing to publish on the Internet given a scale of one to five (one corresponds to *full disclosure*, and five equates to *non disclosure*) also the amount of online time per day (at office and private) is evaluated. At the end of this section there are questions regarding the self-assessment: Estimation of the amount of online data about oneself, and the relevance of reputation caused by these. These questions distinguish between job-related and private.

¹Name, city, street, personal/office email address, instant messenger contact, profession, hobby, phone numbers, and pictures.

Online Search Experiment

This part of the survey is of major importance as it will reveal the online presence of a person according to hits on Google using a given search string. Each participant of the survey is asked to conduct different search queries using the Google search-engine using the following combinations of personal data:

- Surname,
- first name and surname,
- first name, surname, and city,
- first name, surname, and employer/university/school,
- primary office email address,
- primary private email address, and
- primary nickname.

For each query the amount of links (among the first ten returned search results) actually pointing to the person in question, is counted. Additionally, for the last search query (nickname), the appearance of other personal information, e.g., name, address, phone number, on the resulting web pages is reported.

After this Google experiment, the subject is asked if the result meets its expectations, which have been reported before the experiment. In order to take into account whether the name of the subject is very common, rather seldom, or something in between, a database lookup is performed which returns the frequency of the given name, within Germany². Thus, the difficulty in finding persons with a very common name can be normalized. Finally, the subject is requested to query Yasni [93] using the first name and surname in order to report the amount of hits linking to itself.

Social Networks

Since social networks have a major impact on the online presence of a certain person, this part of the survey copes with questions regarding the usage of social networking web sites. Depending on whether a subject is a member of a social networking web site, there are questions on how the different social networks are used:

²Since the survey took place only within Germany a national database provided by <http://www.verwandt.de> [62] was used.

- Job-related or private,
- validity of provided/published data,
- name of the used social network(s), and
- modification of privacy settings.

Personal Pictures on the Internet

Already before social networks established everybody was free to publish arbitrary pictures on the Internet. Anyway, publishing personal pictures became much easier and most of the current social networking web sites even provide a personal gallery or similar features to expose oneself or even publishing pictures of others. In addition there are functions to link photographs to the persons shown on those. The ease of taking a picture and publishing it online leads to a rapidly growing archive of personal histories captured in images. Hence, also these pictures have a major impact on everyone's privacy. Pictures of a certain person might not only give a first impression on its character but also influence its reputation significantly.

In this part of the survey it is evaluated if there are any pictures of the subject on the Internet and also the impression given on the pictures is considered. Here the main objective is to decide whether the pictures provide rather a good or a bad reputation.

3.2.2 Accomplishment

The survey has been accomplished from June 2009 to December 2009. During this period more than 300 people participated. For the evaluation only the first 300 completed forms are considered. All participants have been chosen by chance and not following any certain sampling methodology.

3.2.3 Results

Before turning to the results actually relevant for measuring online presence, we first present some statistics on general information about the subjects. In particular, these are important for reasoning about the explanatory power of the survey: Can the results be applied in general, or are they only valid in this certain case?

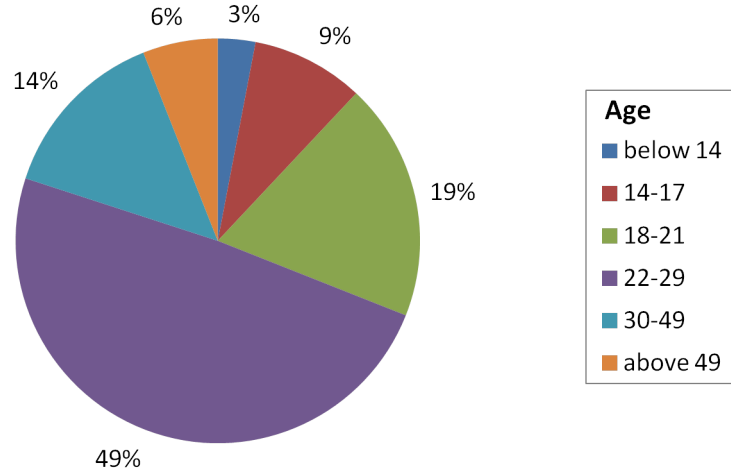


Figure 3.1: Age distribution among survey participants.

General Personal Information

The distribution among male and female participants is 2:1. Most of the participants – almost every second – are within the age range between 22 and 29 years. Furthermore, only 9 subjects are less than 14 years old. Hence, all our results basically apply for persons being 14 years or older. The overall distribution is presented in Figure 3.1.

According to the distribution among different professional categories of the participants, we achieved a pretty good statistical spread. Only the group of IT-related professions is outstanding with about one third of all participants. All remaining participants can be distinguished by 11 professional categories, where one is *other*, containing those, which did not fit into any of the given categories. Figure 3.2 displays the entire result. The majority of people working in IT-related professions, can be explained by the environment the survey has been conducted in.

Exposure of Personal Data in the Internet

Reasoning about factors influencing the online presence of a certain person in the Internet, there are two obvious options: Online time and willingness of exposing personal information. A person never using the Internet might not be present at all in the Internet. Anyhow, it might be possible that even this person is present, e.g., if somebody else publishes information about the person, who has never been online. On the other hand it is possible to

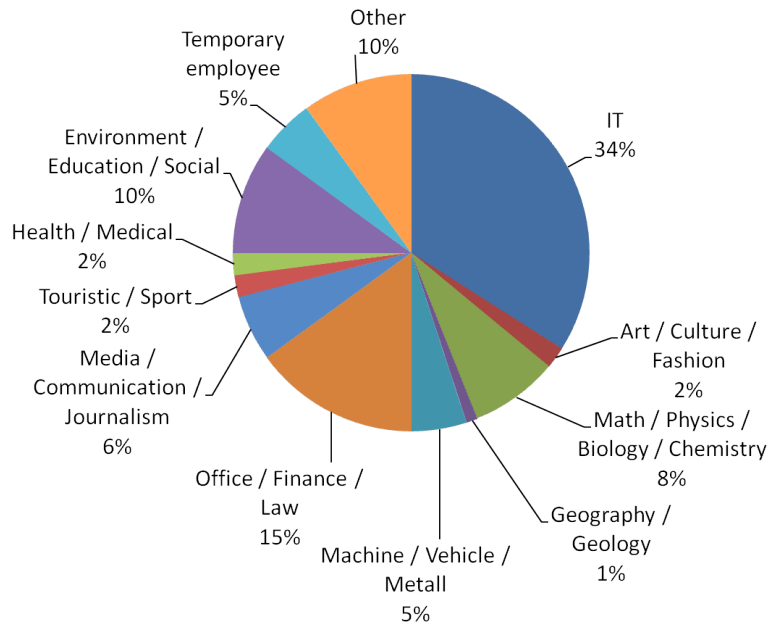


Figure 3.2: Job distribution among survey participants.

spend a lot of time online without actively providing identity attributes at all – apart from information collected in log-files, e.g., search requests, visited web sites, etc.

As these options are most likely to have a noteworthy impact on the online search experiment in the following section of the survey, we first have a look on the answers on these questions provided by the subject. Considering the time people spent online either work-related or private separately, we can observe only 31% spending less than two hours privately online, where 35% spent even more than four hours per day privately online (see Figure 3.3). In turn office-related online time is significantly lower: Only 22% spent more than four hours per day online because of work – most of the work-related online activities, i.e., as claimed by 58% of the subject, takes less than two hours (as shown in Figure 3.4).

The other mentioned aspect of exposing personal data in the Internet is the sensitivity of different personal data as judged by the subjects. Interestingly most people (independent from their age) judge their phone number and street as most sensitive data, even though these data usually can be looked up from a phone book. In turn, other data as for instance job and hobbies are valued to be of lower significance in terms of privacy. Thus, people are

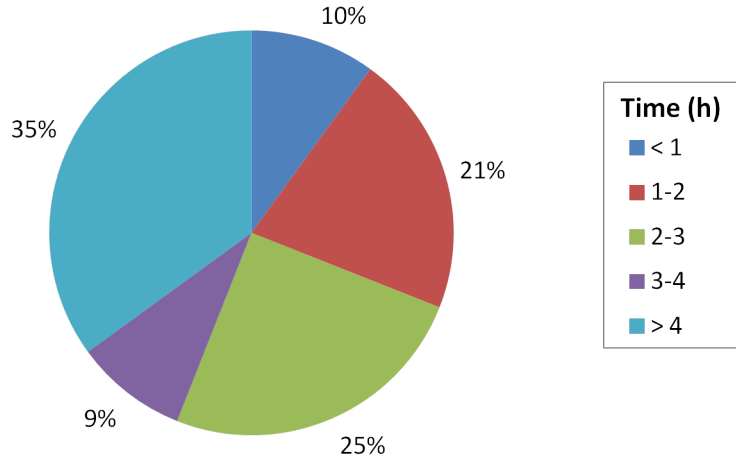


Figure 3.3: Online time spent privately.

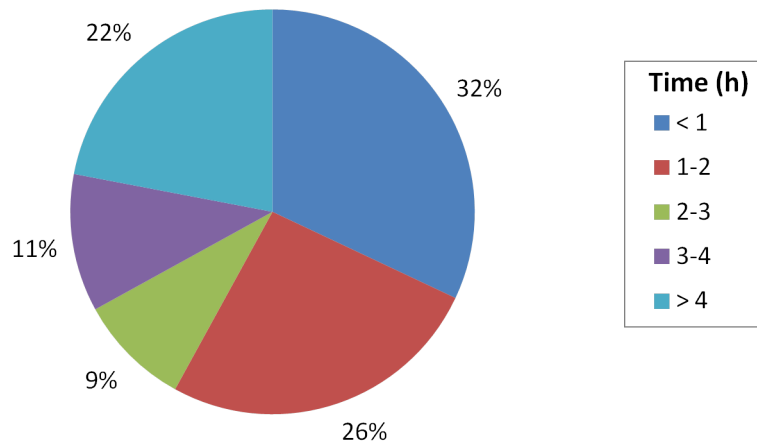


Figure 3.4: Online time spent business related.

not aware of the value of these data if used for profiling and potentially being abused for de-anonymization as presented in recent works, e.g., by Narayanan and Shmatikov [63] or Wondracek, Holz, Kirda, and Kruegel [88].

A general trend is noticeable among the average of all statements given by the participants being 21 years old or younger, comparing to the average of all participants. This can be summarized as follows: The younger the

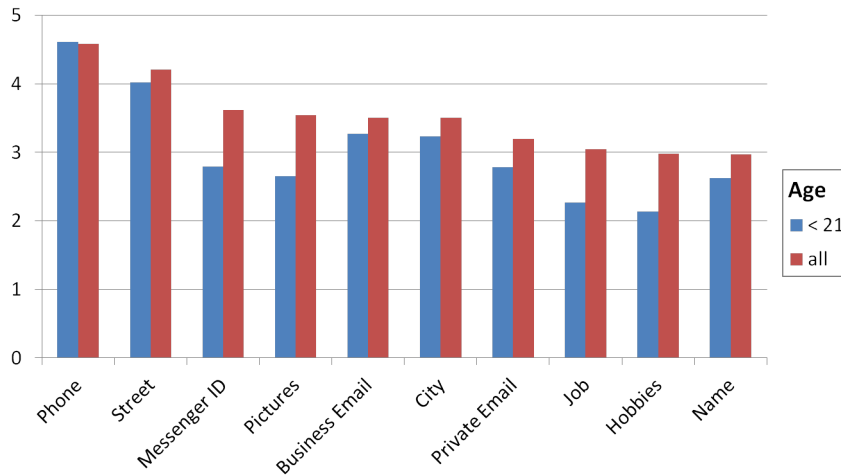


Figure 3.5: Sensitivity to not disclose personal information - comparing all participants with those at age 21 or younger.

person, the lower the estimated sensitivity of certain personal data within the context of the Internet. Only sensitivity in not sharing the phone number is higher among the younger participants.

Figure 3.5 displays the average sensitivity to not disclose certain personal information of all participants, compared to the average among participants being 21 years old or younger.

The last question in this section is targeting the self-assessment regarding the impact of personal data published in the Internet on ones job-related reputation. Asking for the own reputation and the reputation in general the subject where rather less concerned about their own online reputation in respect to their career. In the same time 60% of the participants expect online reputation in general having a major influence on people's career. While this appears rather paradox, it goes with the observation that many people appear to be rather careless regarding their online reputation. Obviously, people tend to think something, which may summarized like this: Online reputation may have a major impact on your career, but who cares about me. The overall statistics on the answers to these questions is presented in Figure 3.6.

Online Search Experiment

In this section of the survey we cope with the most challenging part of the survey, that is measuring the online presence of the subjects. Beside a precise definition on how to actually measure the online presence of a certain person,

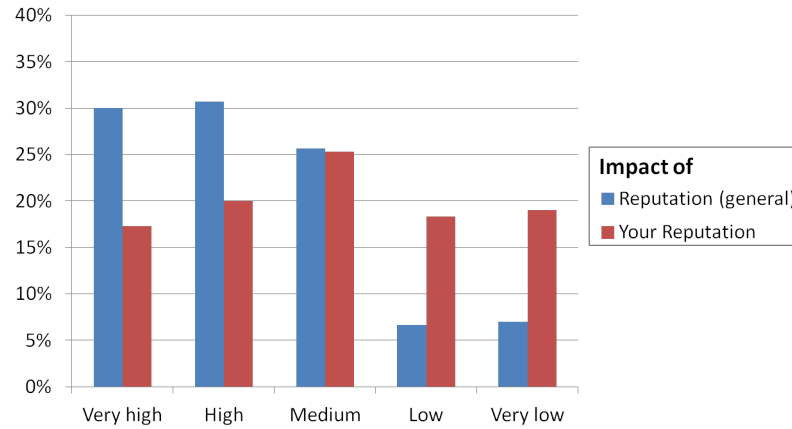


Figure 3.6: Estimated significance of somebody's/your online reputation within business.

we have to consider the privacy of each participant. While analyzing the presence of a certain person within the Internet, we might discover very personal information. Particularly, if we also know private email addresses or nicknames of the person, since these might be chosen with the intention to stay anonymous. Most likely some people would not even provide their nicknames or private email addresses. Thus, the results of the survey would be less reliable. To avoid this, we decided to let the subject analyze there online presence themselves, as described in Section 3.2.1. We are aware of the fact, that we cannot assure all participants have processed this part of the survey in the very same fashion. Still, the instructions are simple in order to render significant deviation in the processing and thus in the results to be pretty unlikely.

For a better understanding of the detailed results, we first have a rather coarse look on the data collected. Therefore, we consider the results by only distinguishing whether a given search returned a hit³ among the first ten search results as returned by Google, or not. The result is pictured in Figure 3.7.

Among the evaluated combinations, first name and surname are most valuable in terms of looking for information about a given person using a search engine. The best results are obtained by using these items in combination with the city or employer. But also the nickname turns out to be useful for gathering personal data, even more than email addresses or just the surname. More than 80% of the subjects published personal information, e.g., hobbies or city, in conjunction with their nickname.

³Result that actually corresponds to the subject.

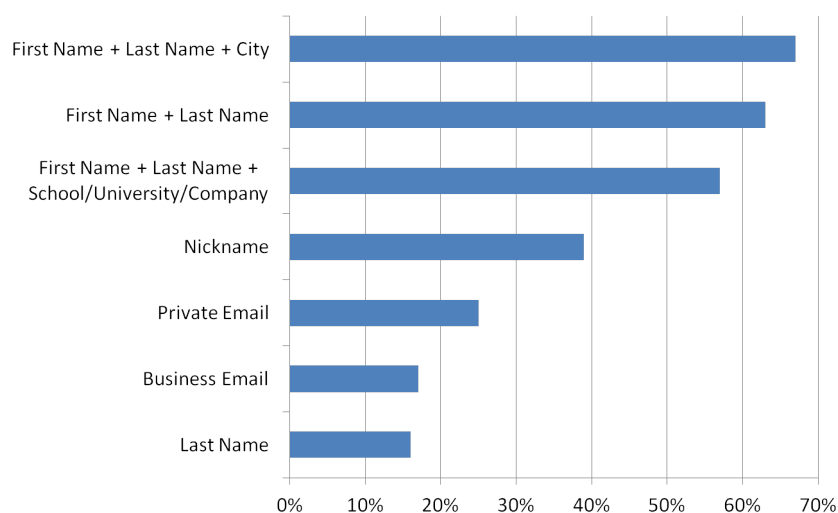


Figure 3.7: Hits among top-ten search results using Google.

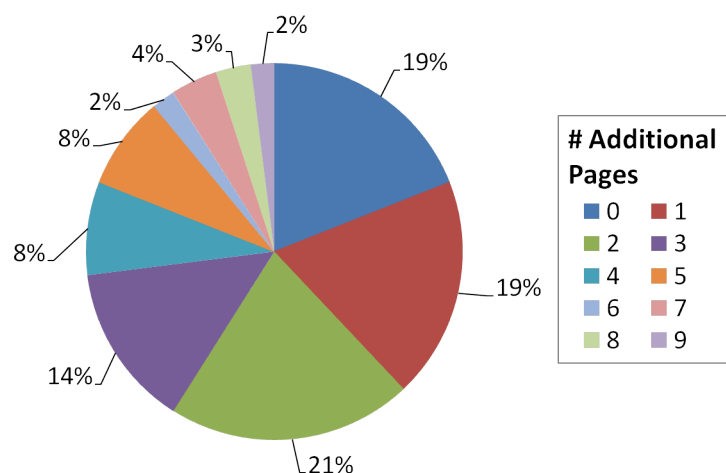


Figure 3.8: Amount of additional pages found searching for the nickname of a user.

Thus, given a nickname and the corresponding first name and surname, we might improve the search results by parsing the data, obtained after searching for the nickname and then reusing these, e.g., city, hobbies, employer, in conjunction with the first name and surname. Figure 3.8 shows the amount of additional pages found when searching for the nickname.

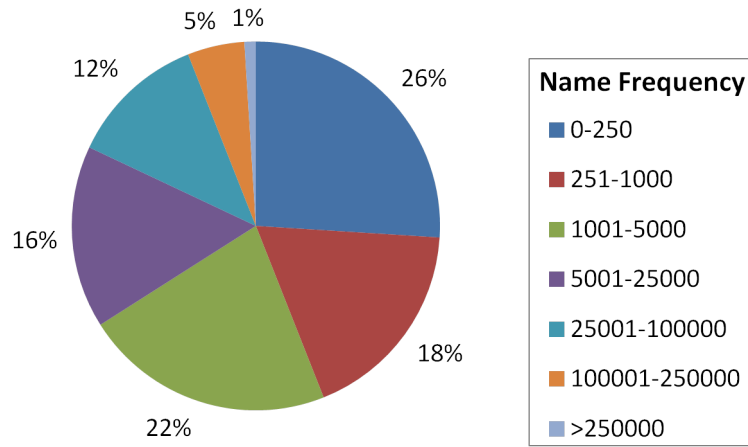


Figure 3.9: National frequency of surname within Germany.

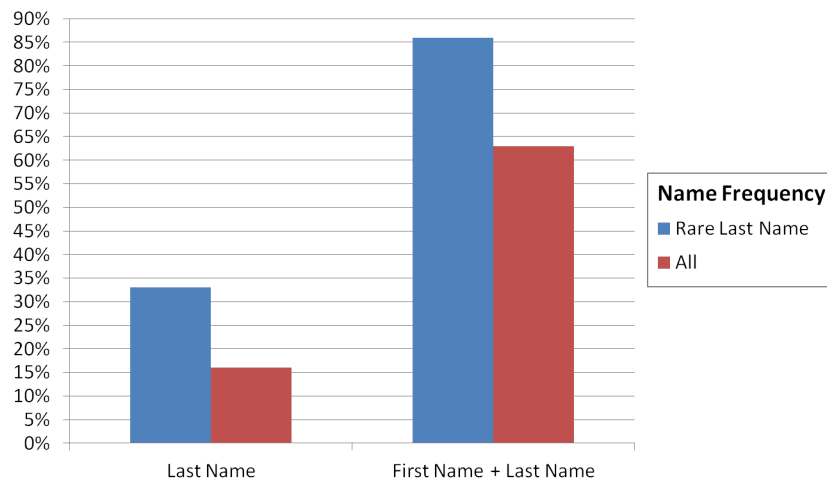


Figure 3.10: Comparing Google hits between all users and users with rare surname.

For a better understanding of the collected data we have also evaluated the national frequency of occurrence of the surnames as provided in Figure 3.9.

This information is important to consider. Comparing the average hit rate between persons with a rather seldom name, i.e., only one to 1000 occurrences within Germany, and all subjects of the survey, resulted in an approximately 20% higher rate as displayed in Figure 3.10, while taking the surname or first name and surname as a query input,

Even though this result is not surprising at all, it is very important for the

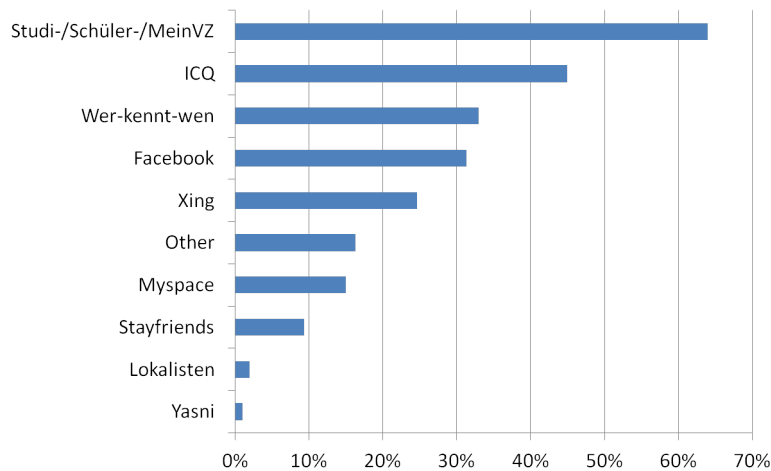


Figure 3.11: Distribution of online social network usage among participants.

research on measuring online presence. Online presence is not only depending on the amount of certain personal data published online but also on the uniqueness of the person in question. While information about a person with a very popular name, e.g., John Cox in the USA or Wolfgang Müller in Germany, is much harder to find, even though the person has published loads of personal information on the web, it might be easy to find information about a person with a rather seldom name, that has published only few information online.

Social Networks

The section of the survey containing questions regarding social networks is only processed if the subject initially states to use certain platforms in the Internet for self-expression. Anyhow, we present some of the results in relation to all subjects whenever it appears to be reasonable according to the context.

Among all subjects 55% use social networks only privately, 4% only work-related, and 21% for both reasons. These numbers support the assumption that the main reason for social network usage is self-portrayal and not business contacts. As expected, this relation depends on the age of the subjects. The younger the subject the lower the usage of business related social networks. Still private usage of social networks is more popular the younger the subjects are. The distribution among the different social networks is depicted in Figure 3.11.

A very positive result in terms of privacy is the ambition of the subjects to

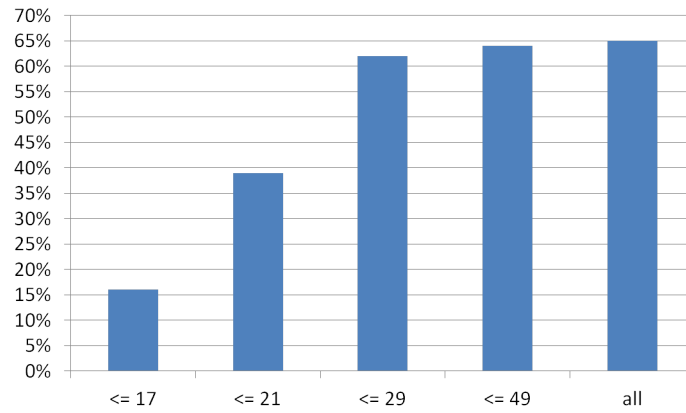


Figure 3.12: Privacy awareness among participants below a given age.

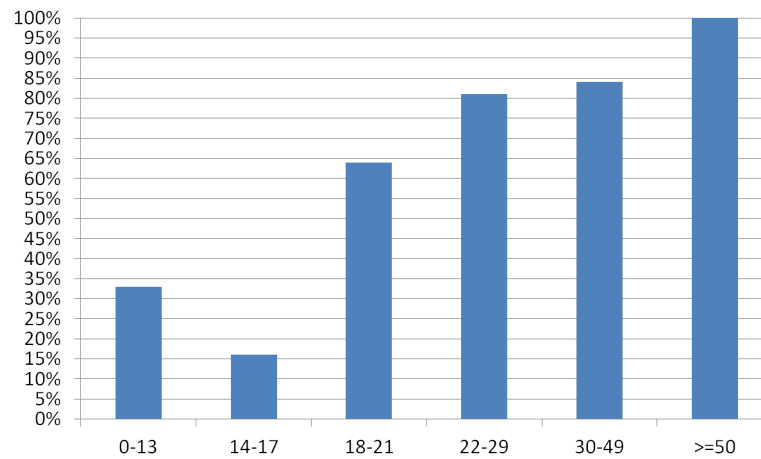


Figure 3.13: Privacy awareness among participants within a certain age range.

adjust privacy settings provided by the platform in use: Three of four users claim to have their privacy settings modified. This supports the observation that 65% of all social network users judge privacy to be important or even very important. While this seems to be quite a good result in terms of privacy awareness, the relative amount decreases a lot among younger users. Considering only subjects being 17 years or younger, only 16% are concerned about their privacy. Considering the relation of privacy aware users regarding the age, we can see in Figure 3.12 and Figure 3.13 that older users are much more concerned about their privacy. One reason for the rapid increase in privacy awareness among users within the age of 18 to 29, might be the career

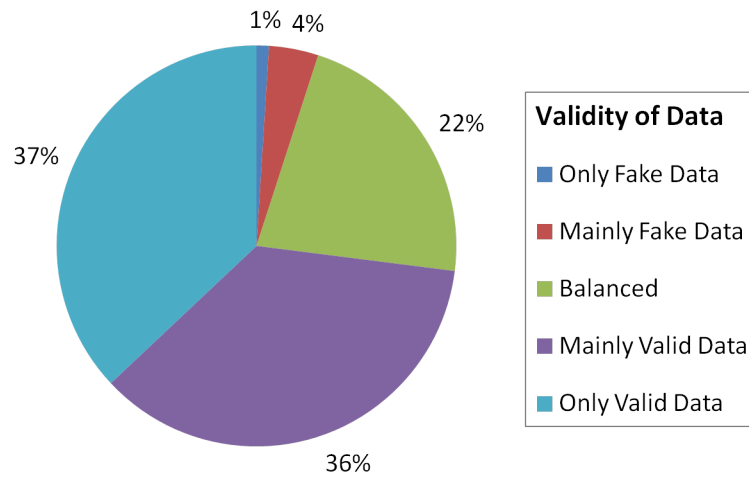


Figure 3.14: Validity of personal information published within online social networks.

entry, which may lead to a change in the reputation people are aiming for. While young people might consider respect and credits in terms of coolness as desirable properties, the older might be aware of the fact that their career is depending on their reputation. In response we can observe a shift of priorities among desirable properties.

Another aspect covered in this section is about validity of the data published within social networks. Do people supply valid data only, or do they also publish data, which do not necessarily correspond to their actual identity? Here, we found 73% to provide mostly valid information about themselves (see Figure 3.14).

Personal Pictures in the Internet

The last part of the survey is considering personal pictures in the Internet. Not being linked to a certain person's name, e.g., by a corresponding filename or caption, they are much harder to find. While this makes gathering pictures of a certain person more difficult for a third party it also renders reputation defense to be a challenge for the person pictured. It is impossible to keep track of all personal pictures published in the Internet. Here, the only chance for preserving privacy is to monitor the pictures, which may be found easily by third parties. Anyhow, there are no public available tools for automated monitoring of personal pictures in the Internet, which can be considered as reliable as necessary. Probably Facebook and Google are leading in terms of effort on implementing such features. For an overview on state of the

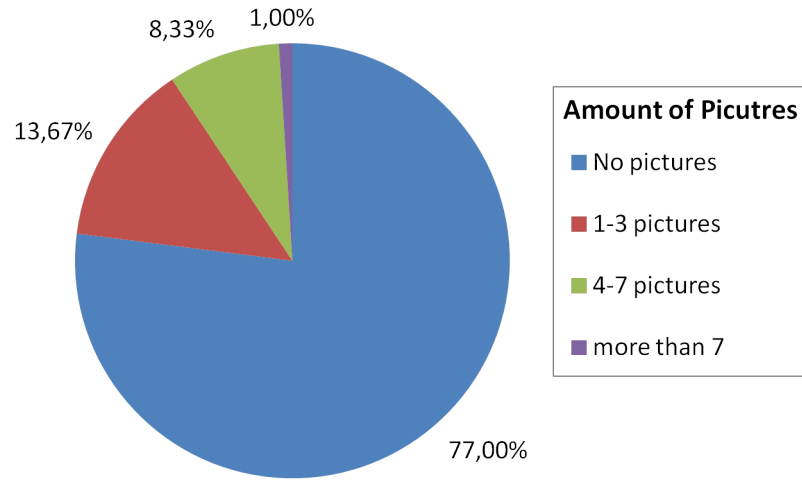


Figure 3.15: Number of hits among the first 18 returned search results using Google image search.

art in this area we refer to the publication *Facial Features Extraction and Applications: A Review* [89] by Wu et al.

In this section the participants are asked to perform a Google image search using their own first name and surname as a query. The results are evaluated by the participants themselves by answering different questions. It turned out that 77% of the participants had no hits among the first 18 search results returned by Google, which were actually linking to pictures displaying themselves. Only 1% of the participants have got more than seven pictures returned showing themselves. The complete result is depicted in Figure 3.15.

However, the willingness of putting personal pictures online is significantly higher within the context of online social network usage, as we will see after the next paragraph.

The following question considers the context of the picture: The participants are asked whether their pictures they have found, rather show them in a work-related context or privately. Accounting only for those participants, who had at least one corresponding picture returned, 66% found private pictures about themselves and nearly 12% even found private pictures they estimated to be potentially harmful for their career.

The last questions of this section, which are also the last questions of the entire survey, refer to personal pictures in social networks. Being asked about picture publishing within social networks 91% among the participants using social networks admitted to have personal pictures online on the cor-

responding web sites, which is 74% among all participants of the survey. According to this result it might appear astonishing that so little results have been reported among the Google image search, but there are many possible explanations. Users of social networks might have adjusted their privacy settings for the pictures. Thus, they can control, if the pictures are listed via Google searches, or not. Furthermore, many users of social networking sites tend to not supply their real name, but rather a pseudonym. Accordingly, the picture might not be returned when querying Google using the first name and surname as search terms. Noteworthy is that social networks seem to turn privacy awareness regarding personal pictures upside down. While three-quarter of all participants are careful with publishing pictures on the Internet in general, there are three-quarter of all participants who become careless when publishing their pictures happens within a social networking platform. Thus, 62% of the participants having more than 19 pictures published on social networking platforms had no hits when accomplishing the Google image search and 94% of the participants using social networks and having no hits on Google image search have published pictures within social networks. Of course, we have to put into perspective that the measurement of online presence using the Google image search is only a coarse measure depending on different factors as for instance reputation of different web sites indexed by Google, or the frequency of occurrence of a certain name. Still, the results show a remarkable tendency. Considering the amount of Google hits in relation to the frequency of occurrence of the different names among the participants, we observe that in average people with no Google hits belong to the category of people who share the same last name with 1001 to 5000 other people in Germany, while people with Google hits rather belong to the group sharing their last name with 251 to 1000 other people. Figure 3.16 gives a detailed breakdown of the corresponding measurements.

3.2.4 Identifying Clusters of Different User Profiles

The results of the survey as presented in the previous section have suggested that online visibility and relevance of identity attributes in online identifiability, i.e., searching for a particular identity given some of its identity attributes, depend on different demographic properties, e.g., frequency of the last name of an identity under consideration.

Therefore, we decided to apply clustering in order to analyze whether dependencies between online visibility and demographic properties can be validated and to better understand the relevant factors telling the clusters apart. The steps taken to prepare the data set and conduct the actual clustering are listed briefly in the following. For further details and explanations

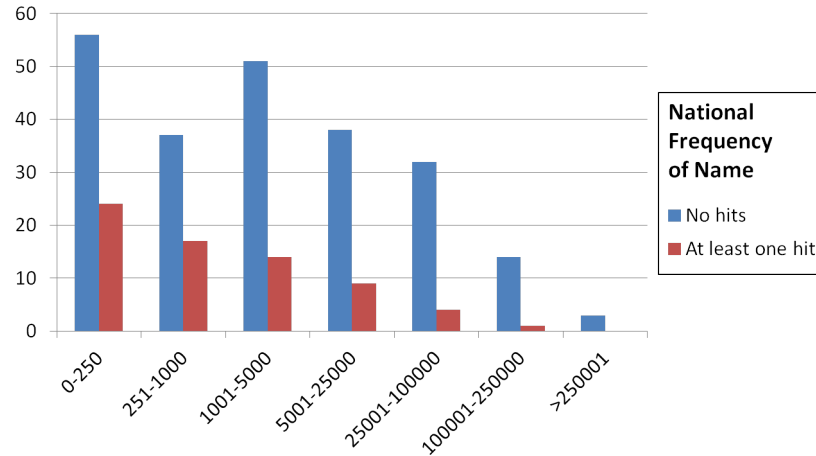


Figure 3.16: Relation between national frequency of surname and Google search results.

on the different methods we refer to Hartigan's *Clustering Algorithms* [43], since detailed description of clustering is out of the scope of this work. Instead, we focus on the results which are presented right after a brief overview on our clustering methodology.

Dataselection/-extraction and -transformation Since the main goal here is to prove whether it is possible to draw conclusions on personal attributes given only identity attributes describing online behavior and online visibility, we first separate the data collected by exactly these criteria. In the following we only analyze the attributes relating to online behavior and visibility to identify clusters as expected. In a second step we will compare the resulting clusters with the demographic attributes, to evaluate our theory.

Clustering is a very complex topic and we do not aim for a complete explanation on the clustering as we performed it. Still we want to sketch what we did you it can be followed by the interested reader, who is experienced in clustering. In case of questions on any details we refer to the corresponding literature.

For normalizing we use *z-transformation* in order to balance variance and arithmetic mean. Next we build a data matrix and correlate the different attributes. This procedure, known as *exploratory factor analysis* [41], eventually helps us to reduce 52 attributes down to eight relevant factors. The reduction has been decided by manual tests and is between the values three (as suggested by applying the *scree-test* [15]) and 14 (as suggested by the Kaiser-Criteria). Since distribution of the attributes among the factors was

rather diverse, we employed the *varimax rotation* approach [49] in order to obtain a well defined loading of the factors by different attributes and thus allow for better interpretation of the results. Following the process as described we eventually found a rotation such as 43 out of 52 attributes load exactly on one of the eight factors. Thus, the remaining nine attributes are not considered for the further processing. The resulting factor loading corresponds to different topics as follows:

- Factor 1: Job related.
- Factor 2: Pictures and Social Networks.
- Factor 3: Willingness to share information.
- Factor 4: Nicknames.
- Factor 5: First and last name related items.
- Factor 6: Reputation.
- Factor 7: Last name related items.
- Factor 8: Private items.

Clustering Given these cluster variables, i.e., the factors as listed above, we use *single-linkage clustering* [46] in order to eliminate outliers and determine the optimal number of clusters. Once the optimal number of clusters is computed, we apply *k-means clustering* [56], providing us with exactly this many (k) clusters, which are as heterogeneous as possible comparing each other with a maximal homogeneity within each cluster. At this point each participant of the survey can be categorized by falling into one of the three clusters.

Results Evaluating the three clusters and the survey participants belonging to each cluster, we notice the following classification.

Cluster 1 Biggest cluster, less personal information, less usage of online social networks and other community sites, private online time rather short, jobs are less IT related, higher age.

Identifying Particulars: Old, low private online time, job not IT-related.

Cluster 2 Many personal information, extensive use of online social networks and community sites, lower age, private online time rather high, job is mainly pupil, student, trainee.

Identifying Particulars: Low age, private online time is high.

Cluster 3 Many job related information, average amount of personal information, extensive use of social network platforms and job-related websites, mid-age, high job-related online time, rather IT-related jobs.

Identifying Particulars: mid-age, high job-related online time, IT-related job

This classification may be used for future surveys and for the design of improved request analysis (compare Section 4.5.3).

3.2.5 Conclusion of the Survey

Comparing the results of our survey, we can confirm trends as noticed by similar surveys conducted before, e.g., by ARD and ZDF in 2009 [6]. Most of these studies we found are not as detailed and do not focus on our particular interest, which is identifying relations between demographic information and online-behavior/-visibility. Conducting our own survey with more than 300 participants helped us to build the knowledge base required for building proof-of-concepts (compare Section 4.6) which implement the dilution concept presented in Chapter 4. Validating our assumption that clustering users according to their online behavior allows for deriving certain demographic particularities, emphasizes the potential of our request analysis as performed in Section 4.5.3. Thus, gathering information about requesters of information in order to decide which information to present to these requesters appears to be reasonable.

3.3 Online Profiling

Digital identities provide valuable information for various stakeholders. The more identity attributes of a certain individual are known, the better its personality can be assessed. Web sites in general are designed to address their visitors as individually as possible to guarantee convenient user experience. Online shops have a particular interest to present to a visitor those articles first which are most attractive and hence will increase the chance of selling. Advertisements are best perceived if they display a message targeting the preferences of their observer. However, not always the purpose of online

profiling is to estimate the best fit (e.g. out of a set of products) for an individual. In case of filling open positions this is vice versa. Here, estimations are aiming to understand whether an individual (the candidate) will fit to the position and the work environment, e.g., team, work conditions, etc.

There are many other examples where information about individuals are of interest. The way to get such information is *online profiling*. In the following we discuss techniques, advantages and short comings of online profiling. The resulting knowledge will be applied in our approach presented in Chapter 4.

3.3.1 Passive and Active Online Profiling

Online profiling can be distinguished in two different categories: *passive* and *active online profiling*. Passive online profiling is mainly applied in online advertising and customized, personal web experience. It is achieved by monitoring the behavior of a web site visitor: What queries does a visitor trigger, which links are visited first, how long does the visitor stay on different sub-sites, etc. These questions and many more can be answered by monitoring the interaction of a visitor with the web site under consideration. Correlating the results with the content displayed allows to conclude on preferences of a given visitor. In turn, the presented content can be further optimized in terms of meeting the expectations and interests of the visitor. The techniques in use and the data analyzed here are also used in our passive polymorphic dilution design (compare Section 5.2). Active online profiling is slightly different and more difficult to automate. Here, personal information regarding a certain individual is collected from all over the Internet. The challenge is to uniquely determine whether a certain identity attribute is relating to the individual in question, or not. No matter how many real identity attributes are known a priori, cataloging further identity attributes is very difficult, particularly in case of virtual identity attributes or real identity attributes which do not correlate with the already known ones. A domain where active online profiling is becoming more prominent is human resource management. While candidates invited for a job interview can prepare and plan how to present them self it is not that easy to control information available on the Internet.

3.3.2 E-Recruitment

In July 2009 the German Federal Ministry of Food, Agriculture and Consumer Protection has conducted a survey (“Umfrage zu Haltung und Ausmaß

der Internetnutzung von Unternehmen zur Vorauswahl bei Personalentscheidungen”) on significance of Internet usage during pre-selection of candidates in human resource management [34]. In this survey 500 companies have participated. The results show that overall 28% of the participating companies use the Internet for human resource decisions. Remarkable here is that among large companies (more than 1000 employees) 46% use the Internet while among smaller companies (less than 100 employees) only 21% employ Internet research. 80% of those companies using the Internet for human resource decisions do so already during the preselection phase and include the results in their decision base, whether to invite particular candidates, or not.

In order to understand E-Recruitment in depth we conducted a similar survey with only four participating companies/organizations. While the scope of this research is less representative comparing to 500 participating companies in the above referred survey, its results reflected the previous numbers accordingly. Furthermore, the smaller scope enabled us to conduct the survey in a very detailed fashion using face to face interviews. One company even let us participate in the application process of two candidates including the interview and the pre and post analysis of the candidates. As an outcome we found the results of the survey conducted by the German Federal Ministry of Food, Agriculture and Consumer Protection confirmed and got further insights of particular attitudes within the selected companies. The provided online profiling was used to review the impression during the face to face interviews. In one case certain information found in the Internet led to an additional interview, to elaborate on the findings. We desist from giving a detailed presentation of the results derived from our survey, since no added value in information could be derived apart from the confirmation of previous surveys and a better understanding of E-Recruitment. Nevertheless, the set of questions used during the interviews can be found in Appendix C. Even though it appears that E-Recruitment is not well defined yet, there are already internal trainings within companies aiming to educate recruiters on how to make use of particular online social networks or the Internet in general in order to find as many additional information on candidates as possible. Unfortunately, such training materials are classified as for internal use only and thus could not be handed out. Asking recruiters about the credibility of personal information available on the Internet, we learned that there is an awareness of possibly misleading information, e.g., placed with the intention to harm a particular individual’s reputation, but in the same time the recruiters confess that it is likely to happen to not have such fake information being distinguishable from real data.

3.3.3 Robustness and Reliability of Online Profiling

After we have seen online profiling being understood as a valuable resource for additional information particularly in human resources management, we started wondering whether such information can be intentionally tampered to successfully change the resulting impression to either become more positive or negative. To elaborate on this question we took the challenge to build up an imaginary online identity from scratch in a way that it cannot be distinguished from a real online identity of an unknown individual. In the following we will present methodology, course of action, and the result of this experiment.

Methodology

Obviously, the success in achieving a trustworthy representation of an imaginary online identity depends on two main factors: Consistency among the identity attributes and the prominence within the Internet. Neither a well planned imaginary identity which is hard to be found in the Internet, nor an inconsistent or unrealistic imaginary identity, which is very prominent in the web will convince a visitor, i.e., requester of such information. Thus, we decided for the methodology as depicted in Figure 3.17.

First, we come up with an personal data sheet as consistent and complete as possible. Since we follow the goal to end up with an extraordinary good reputation, we particularly try to create an educational career including best facilities according to given rankings. After this initial step, we publish the out-coming identity by creating accounts at different online social networks, forums, and other online services, and assemble a homepage accordingly. In order to support consistency of the initial personal data sheet we particularly try to retroactively get the corresponding individual being mentioned for instance on web sites of high-schools we claim to be visited in the past. Next, we review the success of this approach. Wherever, we failed to support certain details mentioned in the personal data sheet by having according statements placed in the Internet, we change certain identity attributes, e.g., the visited university, to some information we found easier to be confirmed by appropriate positioning within the web. Here, for each attribute a trade-off had to be found between lowering the quality of identity attributes in the curriculum vitae (CV) in order to have more sustainable online publications, or favoring the quality of identity attributes and thus looking for alternative publication channels within the web. For example becoming listed as an alumni of certain schools did not work, but instead we were able to establish online friend relationships within in social networks so that our claims were

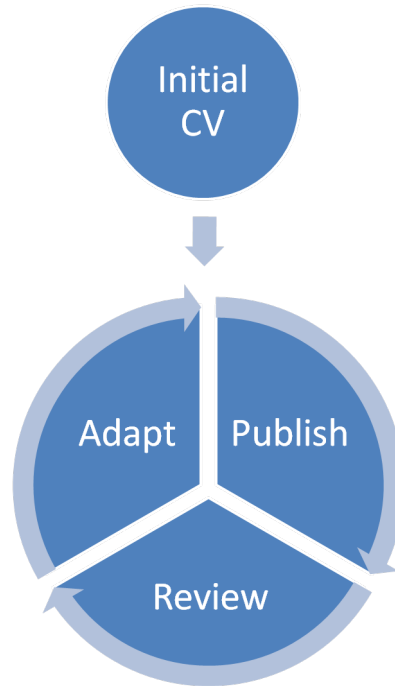


Figure 3.17: Methodology in creating and publishing an imaginary identity on the Internet.

strengthened this way. After this step we again try to publish corresponding partial identities further in the Web, and continue with the next iteration of this circular process.

Course of Action

Following the before discussed methodology, we created accounts at several online social network platforms, published a personal web site, participated actively in different online communities and positioned our imaginary identity as a member of a sports club. We desist from presenting details here, in order to avoid biasing on the online presence of our imaginary identity by the publication of this thesis. Instead we refer the interested reader to search for *Kai Raich*, which is the name of our imaginary individual. At the time of writing we find it astonishing how present the online identity of Kai Raich still is. Considering the fact that the identity has been published mainly in the course of the first half of year 2010, the identity is still very prominent on the Internet. In the following section we will reflect on the results as perceived in 2010, right after several iterations of the above mentioned methodology have been successfully completed.

Results

The results as presented in the following reflect the status of June 2010. For a critical evaluation we impersonate a recruiter looking for further information and confirmation of provided data on a job application based on the personal data sheet assembled before. Searching for *Kai Raich* via a popular search engine provider, we found the first hits pointing to social online network profiles and the personal homepage of Kai Raich. Additional information on student jobs as claimed in the personal data sheet are also visible among the first returned hits. Corresponding online services confirm the educational career. Furthermore, posts in different technical online communities underline the technical interest and engagement as stated in the CV.

Following the links returned during the search engine request for Kai Raich, we are directed to different online social networks. All of them contain only few personal pictures which do not lead to any negative impression. Comments and conversations visible on the personal sub-sites of these online communities are consistent with the amount of friends and do confirm different identity attributes such as memberships in associations, personal interests, and spare time activities.

All in all, we found it impossible to distinguish the online presence of our imaginary individual named *Kai Raich* from any other online identity of a real existing individual. The criteria we found most important to a recruiter (compare Section 3.3.2), e.g., consistency of CV and online presence, engagement, social competence, were perfectly met by the online identity of Kai Raich. Still the available information did not appear to be exaggerated positive, but as natural as of any other person a company would like to welcome as a new hire. This has been validated in a student research project, where one student was asked to take the role of a recruiter and judge on three different CVs. Two students volunteered to provide their CV and the third CV was the one as designed for Kai Raich. The student copying the recruiter was not able to tell the CVs apart with regards to real or virtual, even though she was encouraged to use the Internet in order to verify/refute the credibility of the CVs.

3.3.4 Monitoring Online Reputation

After we have seen how easy it is to create an imaginary identity and position information online to confirm existence of this particular identity we can estimate the risk of manipulating the reputation of a real existing individual to become better or worse. Conclusively, we see the online reputation of any individual at risk. To counter this threat we decided to come up

with a solution monitoring the online reputation of a certain individual. We are aware of online services offering online reputation monitoring or even reputation defending, i.e., improving the online reputation of a given client. However, we find it contradictory to use proprietary services asking you to provide personal information, i.e., identity attributes, to protect your online reputation and hence your privacy. Therefore, we started development of an open source online reputation monitoring framework, which is presented in the remainder of this section.

Online Reputation Monitoring Framework Design

Monitoring of a certain individual's online reputation is a very comprehensive task. Automation of this approach is even harder. Thus, the main focus of our effort is to come up with a modular framework which enables further extensions to be implemented easily. Based on this motivation we suggest an architecture as shown in Figure 3.18. Core of our architecture is a processing unit which provides the main functions. These can be extended by plugins, which interface with the processing unit. The idea is to have one plugin for each supported web resource. A web resource can be for instance a search engine, an online social network, or any other online community or similar web service which can be accessed using a given interface. Besides, a plugin can also be added to extend functions of the core processing unit. Since monitoring is a continuous process a dedicated scheduling unit is implemented. This component will trigger different monitoring tasks and if necessary notify the user, e.g., by sending an email notification. Relevant data is stored in a database by the processing unit. To enable access for third party applications a well defined public interface to the entire system is provided. To ensure privacy for the users of the system strong authentication is required for any interaction with the public interface. The same data interface is also used by an graphical user interface which supports convenient interaction with the framework to all users.

Implementation

For the implementation we used *Java SE 6* and a web interface as a graphical user interface. Thus, the presented solution is independent from the operating system in use. In the following we describe usage of the system starting from the registration of a new user, proceeding with the different processing steps and interactions, and ending with the notification emails send to the corresponding user. This description explains best the functions, which have been implemented already.

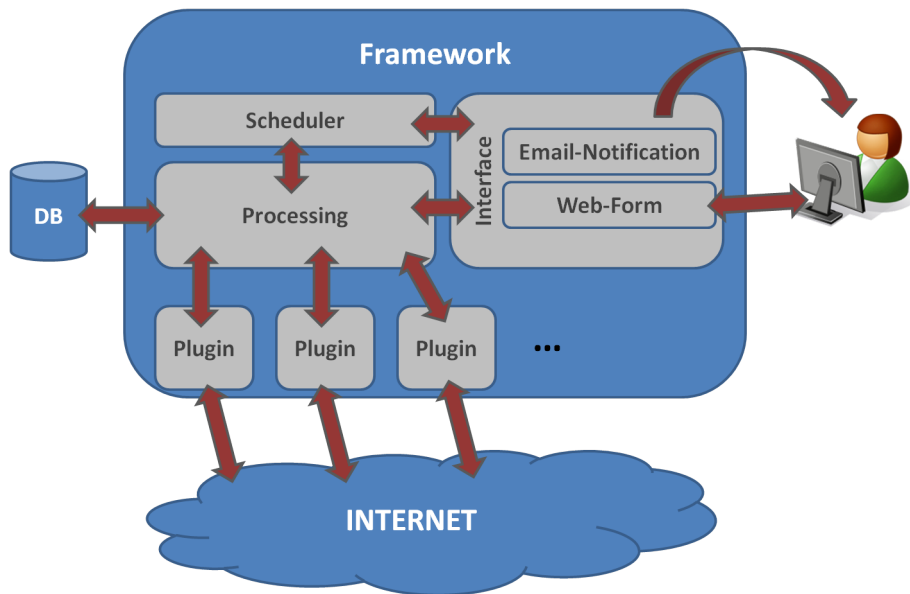


Figure 3.18: Architecture of the Online Reputation Monitoring Framework.

Registration After installation and configuration of the system on a server, which is connected to the Internet, a new account has to be created. For this purpose

- user name,
- password, and
- e-mail address

have to be submitted to the service. Once a user account has been created the system can be initialized.

Initialization Next the system has to be initialized with a customized configuration. There are two options: On the one hand a user can provide its credentials for any online community, which is supported by the framework, i.e., an appropriate plugin exists. This way the system can login to the corresponding service and automatically gather available identity attributes to initialize for the monitoring process. On the other hand a user can manually provide a list of real identity attributes which then will be used during the monitoring in order to distinguish relevant findings from non-related ones, i.e., a relevant finding is any information that refers to the account holder's identity. Any other information is classified as not relevant and hence being

discarded. Optionally, a user may provide account information for the on-line picture service Picasa [37]. If such credentials are provided, additionally processing will be enabled as we will describe in the next paragraph.

Processing In this paragraph we describe all processing steps implemented in the current version of the framework, including a plugin to crawl the online social network service Facebook [30]. Since monitoring is a continuous process a scheduler is employed to trigger crawling processes in configurable time intervals. This way crawling of web resources supported by the existing plugins is subsequently triggered. The gathered information are of two kind: pictures and text. Pictures are further processed using functions provided by Facebook or optionally compared using the Picasa desktop application. The latter mentioned option requires quite some manual interaction and hence, is not convenient⁴. The goal in our picture processing is to identify pictures which either show or somehow relate to the user of our framework. The second kind of information we have to deal with is text. As most of the text, which can be extracted from social online networks is rather short, we haven't found any proven good standard language processing techniques suitable to our scenario. Therefore, we combined different techniques in a best-effort approach. These techniques include

- removing stop words,
- stemming and lemmatization,
- naive Bayes-Theorem,
- Rocchio-Classification, and
- k-nearest-neighbor algorithm.

By combining the different methods our framework can distinguish relevant from irrelevant data. In case of doubt the data is classified as relevant and thus presented to the user for manual decision. Whenever new relevant data is found the user will receive a notification email asking to login and review the latest results. After logging in the user can mark the automatically gathered data to classify these accordingly. Any data which is classified as relevant will impact on future iterations of this processing.

⁴In the past Picasa also offered an online service, which we wanted to employ in the first place. Unfortunately, this service is no longer provided, which enforced us to go for the less handy alternative. A promising alternative may come up in the near future provided by Face.com [31]. During our implementation this service was still in beta-testing and thus could not be considered.

Evaluation

According to the implemented plugin (Facebook), we used a Facebook profile with 80 friends for the evaluation of our system. During the evaluation three different aspects were considered: performance and robustness, processing of pictures, and processing of text.

Looking at the **performance and robustness** of our framework, we found that an initial run, processing all entries on the contact's message board (in case of Facebook this is called *wall*) took approximately four hours. Since subsequent runs are much faster, this duration appears to be acceptable. Also a user commonly does not have to wait for the run to finish but instead will be informed via email whenever the run is complete and if new, relevant entries were found. However, most time consuming in our approach are lookups on external databases publicly available via web services. For instance the service Thesaurus [21] allows for maximal 60 requests per minute. Thesaurus, a synonym database is used for the stemming of identity attributes. The results for each lookup are stored in the local database, so that subsequent lookups are much shorter in time. In our initial run querying external services resulted in 9300 additional entries in our local database. During the clustering of gathered data we experience system instabilities due to a lack of memory. After increasing available main memory from initially 128MB to 1024MB this issue was solved. Additionally we limited the amount of considered message posts to 300 per run and contact.

Apparently, **processing of pictures** is not mature enough yet, for our purposes. We observed major problems when trying to compare images with different resolution, different brightness, or taken from different perspectives. There are promising projects like for instance *Face.com* [31] but these were in a beta-state during the time of this writing. However, this might be a particular problem related to Facebook and its, or its users attitude. Most of the pictures in Facebook are rather snapshots taken in parties or at various different places. Other online social networks as for example Xing [90] or LinkedIn [55], usually have rather professional profile pictures as for instance found in a curriculum vitae. Using state of the art methodology such pictures are much easier to compare as our experiments have shown. Therefore, we would like to encourage reconsidering picture processing in case of further plugins being developed for different online social network sites. Also a review of the state of the art [89] in the near future might suggest novel approaches to be implemented in our Facebook plugin.

In **text processing** we implemented three different techniques using: *naive Bayes-Theorem*, *Rocchio-Classification*, and *k-nearest-neighbor algorithm* [43]. During the first runs of our framework the approach using naive

Bayes-Theorem is rather slow and not very precise, i.e., it judges to many entries as irrelevant, even though they are not. Rocchio-Classification and k-nearest neighbor algorithm both are much faster on small data sets, but require an initial classified dataset in order to operate as expected. Since also the Rocchio-Classification falsely dropped many entries as irrelevant, the k-nearest neighbor algorithm was found to be most reliable. Nevertheless, the text processing methods we tested are rather meant for longer texts like emails, articles, etc. We believe that further research in this area conducted by natural language processing experts is promising to increase the results significantly.

We desist from giving overall numbers in terms of false positives/negatives, since these numbers could not be raised in a empiric way. Either such data have to be created manually and thus might be biased by the intention of this research, or existing datasets, e.g., real account information, are used. In the latter case, all posts would need to be verified manually, which is already for one account a very complex task and for empiric results we would at least need manual classification for 100 accounts.

Outlook

Complementary, we worked on a different approach in monitoring the own online reputation implemented as a Mozilla Firefox [61] add-on [70]. This add-on allows the user to store personal data locally. These data will be used in periodic, automated queries sent to search engines in the background while the user is browsing the Internet. Returned search results are monitored and whenever a new hit is found the corresponding link will be shown to the user. Then the user can review the corresponding results and validate whether the returned results, i.e., hyper links, point to a web site related to the user, or not. All search results along with the result of the user's manual review are stored in a local database. Results which were already reviewed by the user will not be shown in subsequent background queries. Since the entire processing, except from the validation, is happening in the background the required user interaction is kept as minimal as possible. The work flow is shown in Figure 3.19. By now, the add-on works independently from any other solution except from public available search engines.

However, during the design of the add-on it was planned to have it interfacing with the framework presented in Section 3.3.4. Thus, we would like to direct future work to integrate both approaches so that a user conveniently can monitor its online reputation from within the browser, while the processing is outsourced to an external server. The server can be provided as a public service or for personal use only. We want to stress, that we strongly

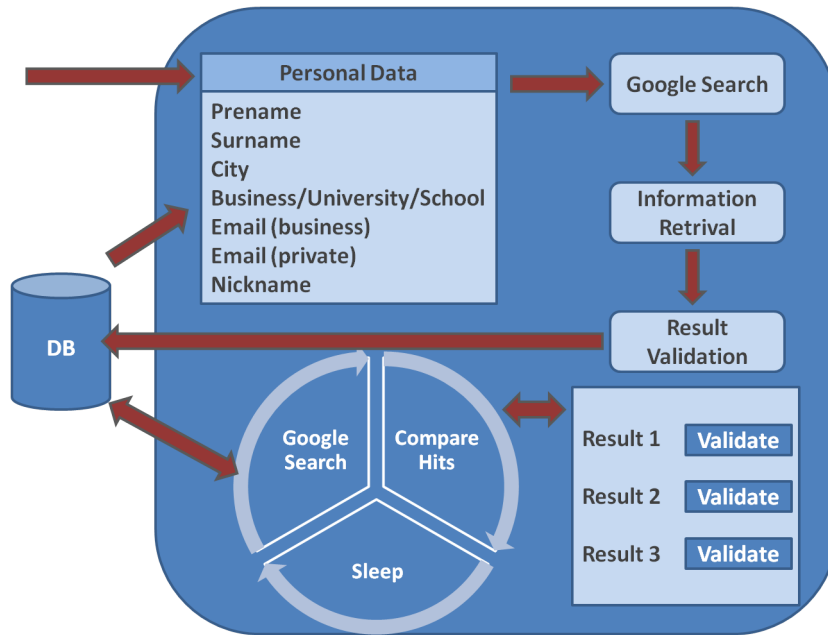


Figure 3.19: Work flow of online reputation monitoring Firefox add-on.

support privacy related services being enabled to be set up and operated on a dedicated server owned by the user. Central service providers, might be an alternative for users, who do not want to maintain their own server. However preserving full control over personal data translates straight forward to enable maintenance of privacy supporting systems to be done entirely by the user of such system.

3.4 Conclusion

After presenting compilation, accomplishment, and results of a survey we conducted to understand relations between personal details, behavior, and online visibility. We discussed how personal details and behavior can be monitored, where we distinguish between active and passive online profiling. Furthermore, we presented briefly the results of another survey evaluating the impact of online visibility on decisions during e-recruitment. The risks in trusting the online visibility of an online identity we illustrated by conducting a case study: by creating an imaginary identity that we position via a crafted online visibility, we show how easy it is to manipulate digital, public available information in a way that those still appear to be consistent. Last in this chapter we presented a framework along with a Firefox plugin, both available

as working proof-of-concepts, which are our solutions to online monitoring without trusting (and handing over personal data to) an online service like for instance Reputation Defender [71].

Chapter 4

DILUTION

4.1 Introduction

In this chapter we present the main contribution of our work. After extending the set of definitions as provided in Section 2.2, particularly by formally introducing the term *dilution*, we will review history with regards to phenomena that can be understood as dilution or at least being close to our definition. Next, we sketch our idea of dilution in brief in Section 4.4, before we elaborate on the design concepts and implementation pillars in Section 4.5. The value of our proposal is proven in two different applications in Section 4.6, which we implemented as fully working prototypes. Assessment of the entire concept by looking at both applications is presented in Section 4.7.

4.2 Definition

The most prominent approach in keeping certain information secret is to not publish such information. Another, not that obvious approach is to not keep such information secret at all, but instead publishing it along with a massive amount of further information. This is what we refer to as dilution. Given natural limitations in time and resources, reaching a critical amount of dilution results in no advantage over not having any information at all. The assumption here is to not allow any measure of the quality of those data.

Following the dilution methodology there are two different approaches we distinguish. According definitions will be given in the following.

Definition 4.1 *In **Mimetic Dilution** additional information is published beside the actual information to protect. As a result different information can be perceived in the same time (in parallel) what renders it arbitrarily*

hard (depending on the amount of additional data) to identify the relevant information.

Definition 4.2 *In **Polymorphic Dilution** data is not published statically, but dynamically. Thus, the data may differ every time it is requested. As a result there is only one kind of data at a certain time (request). Over time (repeated requests) this data changes. Depending on the amount of data available in the back-end identifying the relevant data is rendered arbitrarily hard.*

Next we have a look at recent history, to see whether we can identify phenomena at least being close to a fit to one of the before given definitions.

4.3 History Review

No need for going far back in history there are plenty of examples relating to dilution. However, those are usually motivated by a different purpose. In the following we will highlight a few scenarios where dilution is involved to support better understanding our approach.

Online Banking Fraud In the Internet there are loads of web sites copying common online banking portals (also known as *phishing sites*). Their only purpose is to trick innocent victims in order to steal money. The main reason for the success of this kind of fraud is that people cannot easily distinguish a malicious online banking site from a legitimate one. In turn, an experienced user knows how to determine the location of the web site and hence can recognize such a malicious web site being hosted on a web space different from the legitimate provider.

Decoy Port Scan Port scanning is a technique often part of the reconnaissance phase when an attacker tries to find vulnerable services or not well protected target systems. For this reason an attacker can send packets to all machines connected to the same network in order to trigger responses and thus determine potential targets. An attentive user may investigate the traffic and identify the source IP address which is part of each data packet. This will lead right to the issuer of the port scan. In a decoy scan the attacker transmits additional packets with spoofed source IP addresses. Therefore, it is impossible for the machine under attack to distinguish the actual source IP address from the once being spoofed.

Production Honeypots Some production environments, which are interconnected with a computer network, are protected by so called production honeypots. A production honeypot is commonly placed within a network where productive systems offer services via network access. While legitimate clients know which systems are actually offering such services an attacker usually has to gather such information by try-and-error. Here, a production honeypot behaves like the productive system and therefore cannot be distinguished from a network perspective - same services are offered on the same ports. Hereby, the existence of the productive system is diluted by the coexisting honeypots.

All the above given examples correspond to the **mimetic dilution** approach. Even though the intention differs the impact is the same: a benign online banking web site cannot be distinguished from the surrounding fraud web sites, the original source IP address cannot be distinguished from the spoofed ones, the production system cannot be distinguished from the additionally deployed production honeypots. Next we turn to a very common example which we can relate to **polymorphic dilution**.

Virtual Hosts The technique of virtual hosts allows different web sites being co-located on the same web server using only one IP address. Given a set of different domains all resolving to the same IP address, i.e., the one of the previous mentioned web server, all clients intending to access a web site linked by one of the domains will be connecting to the same web server instance hosted on the same machine, i.e., the same IP address. Looking behind the scene any web browser application will transmit the domain name originally used to access the web site to the server without the awareness of the user interacting with the web browser. The web server again will use this additional information to show different web sites depending on the actual value of the *host* header, i.e., the domain name entered as part of the URL into the browser. As a result the same web server instance provides different content depending on some additional information not visible to the user.

Web Profiling In web profiling all available information related to a certain user visiting web sites is collected and analyzed. This includes, web sites visited, duration of visit, interaction with the web site, etc. As a result user profiles can be derived which provide a useful input when it comes to customized/personalized advertisements. In practice this means different users get to see different advertisements, depending on meta data related to them and collected silently in the background.

The previous discussion of different dilution examples should give a fairly good understanding on different applications provided by this methodology. However, all given examples do not address privacy. In contrast the web profiling example even subverts users' privacy. We see that both kinds of dilution, i.e., mimetic and polymorphic dilution, are already applied in practice. Still this methodology has not been considered as a privacy enhancing technology yet – at least not to our knowledge. In the next section we will present our idea on how to employ polymorphic dilution as a privacy enhancing technology.

4.4 The Idea

In physical life individuals usually share different personal information, i.e., identity attributes, depending on the addressee. This is not only about secrecy, e.g., to keep certain information confidential, but also a matter of common interests. The very same person might share entirely different personal information with colleagues at work than with her family. Of course, there are also individuals not distinguishing between different addressees, but sharing any information with everybody, or not sharing information at all or keeping it at a minimal level, i.e., sharing only necessary information. However, privacy is not about secrecy, but about the right to decide which information to share with whom (compare Section 2.5). Reviewing the state of the art in applications of privacy enhancing technologies (e.g. Facebook [30], LinkedIn [55], etc.) we found that most approaches enable a decision on which information to share. Some of these techniques also allow for static decisions depending on the addressee. Therefore, the addressees are cataloged into generally three different categories: friends, friends of friends, and unknown.

Within this work our goal is to present an alternative and novel approach on how to enable privacy in a dynamic fashion as given in physical life. Following this idea we derive a model from real life interactions and transfer it to the digital world, i.e., the Internet, employing decision algorithms and dynamic web techniques. The general assumptions are:

1. The more a person knows about a particular individual, the more information can be shared.
2. The kind of information which is known to a person about a particular individual allows for concluding on which information to be shared.

Assumption 1 reflects on the fact that the amount of information being shared heavily depends on the level of trust between two communication peers: The

more I trust a certain person, the more information I share. Assumption 2 is taking the quality (or semantic) of the already shared knowledge into account: Somebody I have shown pictures of my last vacation is most likely considered to be trustworthy enough to also see pictures of other vacations I have taken before. While on a high-level view dilution is as simple as enriching a set of real identity attributes with a massive amount of virtual identity attributes the main challenge is *re-purification*: Given a request for personal information we have to decide the *amount* and the *kind* of real identity attributes to share. As an input we process the amount, the quality, and the type of information already included in the request. In response we return a set of identity attributes, i.e., a partial identity, assembled from real and virtual identity attributes, which varies in relation, i.e., amount of real and virtual identity attributes, and semantic, e.g., personal identity attributes, business identity attributes, etc.

As a result we provide a polymorphic diluted partial identity in response to every request.

4.5 Design Concept and Implementation Pillars

In order to implement polymorphic dilution as a privacy measures there are five different phases:

1. Initialization,
2. publication of diluted identity,
3. request analysis,
4. partial identity composition, and
5. delivery of the composed partial identity.

4.5.1 Initialization

During the initialization both, real and virtual identity attributes have to be entered to the system.

Real Identity Attributes During the initialization phase a configurable set of real identity attributes has to be entered to the system by the user, i.e., the identity owner. All provided information are stored within a database along with additional information gathered using different web services like for instance Thesaurus [21]. These additional information will allow us later to also link related words, expressions, job descriptions, etc. to the real identity attributes provided by the user, even though there is no exact match.

Virtual Identity Attributes In order to provide partial identities which vary in amount and quality (in terms of real or virtual identity attributes) of the presented attributes we need additional identity attributes, which can be used during the partial identity composition phase (described below). In a online social network scenario this is rather easy since all real identity attributes related to other participants than the one under consideration can serve as virtual identity attributes for it. Here, the level of privacy protection depends on the amount of participants and hence the amount of virtual identity attributes, of course. In any other scenario we have two options:

1. Manually provide virtual identities.
2. Automated gathering of virtual identity attributes.

While the first option obviously requires a lot of manual effort it comes along with the advantage of having the decision which virtual identity attributes to use. Beyond that one can also compose entire virtual partial identities, which can be used later as defined before hand, e.g., either as manually entered or dynamically composing virtual partial identities using identity attributes from different identities. Particularly, a basic identity can be defined here, which contains a basic set of real identity attributes. This can be used within the partial identity composition phase if the requester can not be determined clearly.

The second choice is much more convenient since public available information can be gathered from the Internet and then being reused during the partial identity composition phase. The drawback here is the lack of control on the quality, i.e., semantic, of the identity attributes being included in composed partial identities. The risk of creating virtual partial identities with bad reputation, i.e., including negative identity attributes, has to be taken into account. For further research on this we hand over to the linguist professionals, since this is beyond the scope of this work. Instead we rather focus on technical aspects in order to implement our idea.

4.5.2 Publication of Diluted Identity

Reaching out for visibility we need to define a superset of virtual partial identities (including the real partial identity) which is entirely published on the Internet. For this purpose we create a dynamic web site which returns the before composed superset whenever a well known search engine is crawling our web site (in search engine optimization this method is also known as *cloaking*, i.e., presenting different content to search engine spiders than to regular users). Then all identity attributes contained in the superset are presented in a mixed way so distinguishing real and virtual identity attributes is impossible. Search engine crawlers are identified by their **USER-AGENT** string. Of course, this header could be spoofed by a potential attacker who would like to see all available information. Still, this won't bypass the protection provided by our dilution approach, due to the massive amount of useless information the real identity attributes are mixed with and the impossibility to distinguish real from virtual identity attributes. The user requesting all these information may only guess or know real identity attributes, given a certain knowledge about the individual who owns the corresponding identity. This again is aligned to our goal: Sharing varying amount and kind of information depending on the addressee's a priori knowledge.

Following the steps as described before we assure to have search engines being aware of the online presentation of our real partial identity. Nevertheless, the search engine knows also about all the virtual identity attributes we have added and can not tell real and non-real identity attributes apart.

4.5.3 Request Analysis

Considering the overall goal to present different information - namely partial identities - to different people requesting these information, we have two major roles in this interaction: The identity owner who publishes/provides personal information (P) and the requester (R), who is asking to receive these information. The requester can be for example a human, a web crawler, an information harvester, or whatever else might be interested in receiving personal information. The purpose of a request can be manifold. While such a request can be motivated by simply establishing contact to the identity under investigation, it may serve as a mean to collect additional information previously unknown to the requester. For the request analysis we present two different approaches *passive request analysis* and *active request analysis*, both of which are presented in the following.

Passive Request Analysis

The passive method of request analysis is most convenient from a users perspective. It does not require any interaction since it is completely transparent running behind the scenes. As an input meta communication between the web server hosting the personal information and the web browser (or client program with similar function) is used. By now the following information is used (further data might be included in the future):

- Values/variables within the URL
- IP address
- Referer header
- User-Agent header
- Cookie header

Next we explain the use of each value. Drawbacks of this approach will be discussed jointly at the end of this section.

Values/Variables within the URL Variables transmitted as a part of the URL can be processed in various ways. Picturing an invitation scenario P may send out links to different R s containing a variable, which holds an identifier unique for each R . When the web server is receiving a request containing such an identifier within the URL the request can be linked to a particular R easily. In the same way variables within URLs can be used to link the web site containing personal information from within other web sites. Again R s following this link will perceive a predefined view.

IP Address Given the IP-address of R further analysis can be performed. Using so called geo-IP-lookup services the geographical origin of the request can be determined. Hence, we can not only apply black-, gray-, and white-listing approaches, i.e., definition of IP address pools, which can see no, some, or all real identity attributes, respectively, but also determine from which geographical location the request origins. To better understand the impact of such knowledge we display the following scenario: A person living in one city and working in another might be able to fully distinguish requests related to its profession and requests related to its private life entirely by having the geoIP information at hand.

Table 4.1: URLs of different search engines using the query *christian gorecki computer science*.

Bing [59]	<code>http://www.bing.com/search?q=christian+gorecki+computer+science&qs=n&form=QBRE&pq=christian+gorecki+computer+science&sc=0-17&sp=-1&sk=</code>
Google [36]	<code>http://www.google.com/search?client=ubuntu&channel=fs&q=christian+gorecki+computer+science&ie=utf-8&oe=utf-8</code>
Yahoo [92]	<code>http://de.search.yahoo.com/search;_ylt=A7x9QX4IpVBPY0cA.1EyCQx.?p=christian%20gorecki%20computer%20science&fr=404&fr2=sfp</code>
Altavista [91]	<code>http://us.yhs4.search.yahoo.com/yhs/search?p=christian+gorecki+computer+science&fr=altavista&fr2=sfp&iscqry=</code>

Referer Header The **Referer** header holds the URL of the web site from where R has been redirected - commonly by following a hyper reference - to the web site of P . Thus, this information can be used to compose the partial identity depending on the domain, e.g., Facebook or LinkedIn, R has visited before. Particularly, in the case of Facebook or LinkedIn this information turns out to be valuable to distinguish business and private relationships. However, this data field can even contain more valuable information if R has used a search engine query to find a link pointing to P 's web site and then has followed that link. In this case the **Referer** does not only include information on the search engine that has been used, but additionally comes along with the search terms being entered during the search query. These are given as values of variables within the URL (compare Paragraph *Values/Variables within the URL* above). Some examples can be found in Table 4.1.

Extracting the search query (here: "*christian gorecki computer science*") is straight forward for each search engine URL. Therefore, our request analysis provides us with the search terms being used to find P 's web site and hence gives a hint on what R already knows about P .

User-Agent Header The **User-Agent** header can be used to extract information on the browser used by R . While there is no added value for our approach by knowing whether R is using Microsoft's Internet Explorer or Mozilla's Firefox, this value allows us to detect search engine requests. Table 4.2 shows relevant information which can be extracted from the User-

Table 4.2: User-Agent substrings to identify search engine bots/crawlers (03/02/2012).

Google	Googlebot/2.1
Bing	bingbot/2.0
Yahoo	Yahoo! Slurp
Altavista	Scooter

Agent header.

This information is relevant to enable indexing of both real and virtual identity attributes as described in Section 4.5.2.

Cookie Header If a **Cookie** header is present the contained cookie is analyzed. Cookies can be used to recognize a requester who has visited P 's web site before. This way the same content can be shown on independent requests by the same person - assuming the reuse of the same client system and the same browser. However, the content might change despite the presence of a cookie, for instance if the request was preceded by a refined search engine query: A requester searching for "*Christian Gorecki*" might doubt the resulting page being a mix of real and non-real identity attributes and consequently do another query using a refined search, e.g., "*Christian Gorecki computer science*".

Limitations

Even though passive request analysis is very convenient in a sense that it does not require any additional user interaction by R , it also comes at a price in terms of reliability. Assuming a well-behaving R passive request analysis might be not as comprehensive as it could be if R would be aware of the processing behind the scene. In the latter case R might consider to use extended search queries right away instead of first trying to get a hit with less information provided.

Another issue comes along with the capability of spoofing, modifying, or deleting header information: Referer, User-Agent, Cookies, or other variables within the URL might be arbitrarily changed or deleted. Thus all these information are not reliable at all. The question becoming most important though is: Can an attacker by-pass dilution as a privacy measure by modifying, adding, or deleting some or all of the meta information used during our passive request analysis? The answer is: No.

Changing values of variables within the URL might lead to a different partial identity being presented, but will not increase certainty on which

identity attributes are real. *IP address spoofing* does not work in such scenarios since the attacker would not receive any response this way. *Altering the Referer header* will only improve the returned partial identity in terms of credibility if the search engine query relevant values are refined according to the real identity attributes of P . This is intended and can also be achieved by changing the search query in the first place. Any other modification of the Referer including changes to variables within the provided URL will only result in receiving a different partial identity, without any certainty on the amount of real identity attributes. The *User-Agent* header will only have impact if it is changed to one corresponding to a known search engine. Then all available identity attributes will be returned without any indication of credibility. *Cookies* may be protected by encryption. Nevertheless, changing a cookie will not allow for spotting real identity attributes, but again will only change the partial identity being returned.

In the above paragraph we can already see that by-passing our approach is at least not easy. An in-depth evaluation of the robustness of our approach is presented in Section 5.

Active Request Analysis

In the previous section we presented our passive request analysis working in an entirely transparent manner. We have seen usability coming at its price that is low reliability on the meta information used as an input. While this does not subvert privacy it still lowers quality of data, which are returned and hence might result in less real identity attributes being presented than intended. To overcome this drawback we next present an alternative, active request analysis. Here, the requester needs to interact with our web site and will receive a partial identity accordingly. This approach can either complement the passive request analysis, e.g., when too less meta information are available, or can be substitute for the passive methodology.

Instead of relying on little information transmitted in the background (passive request analysis) we now introduce some interaction with R . No matter how R happens to request content from P 's web site, a landing page will be interposed showing a configurable set of questions. Of course, these questions relate to personal information, i.e., real identity attributes, of P . They may vary from asking phone numbers, hobbies, profession, favorite holiday location, etc. By answering some or all of the questions R can prove its knowledge in personal information about P . Answering the questions can vary between choosing from a drop-down menu and entering free text input fields. Once the answers are submitted a decision algorithm is triggered as part of the *partial identity composition* phase, which will be presented in the

following section.

4.5.4 Partial Identity Composition

During this phase results of the Request Analysis phase are used to compose a partial identity to be shown to R . The composition is depending on further conditions, which are described within this section. Accordingly, the final output, i.e., the partial identity being presented, can be either a real partial identity or a virtual partial identity. In the latter case the amount of real identity attributes may vary from no real identity attributes, over some real identity attributes, up to mainly real identity attributes. A special case here is posed by the basic identity (as mentioned in Section 4.5.1) which is a predefined minimal set of real identity attributes, for instance providing just enough information to enable an initial contact. Next, we have a look at the conditions which are evaluated in order to determine which of the before mentioned profile information will be returned and how many real identity attributes will be contained.

The overall algorithm implements a decision tree as shown in Figure 4.1. Here we present the decision tree as it looks like in case of combining passive and active request analysis. Adaption for either passive or active request analysis only is straight forward.

First the **IP address** of R will be processed. If the IP address is black listed a fake identity will be used to compose a profile web site accordingly. If the IP address is contained in the whitelist, real identity attributes will be used for the partial identity that is returned. Actually, we also foresee the option to have predefined profiles being returned for certain “white-listed” IP addresses. Thinking of scenarios that include, living in a certain area, working in another area, or applying for a job in a certain location, we preferred this approach rather than having the IP address evaluation being included in a ranking as we will do for some of the other input values. A general remark here is that black- and white-listing IP addresses also includes geoIP location featured information, like cities, regions, or countries.

In a second step the **User-Agent** is compared with a list of known search engine user agents. In case of a match cloaking technique ¹ is used to present all available identity attributes, i.e., both virtual and real.

If the User-Agent string is not known to be used by a search engine spider the **Referer** is analyzed. If the profile page was linked from a web site, which is white-listed, either the real or a predefined profile will be returned. If the

¹Here, a technique to show a different content to a search engine crawler than to a regular user.

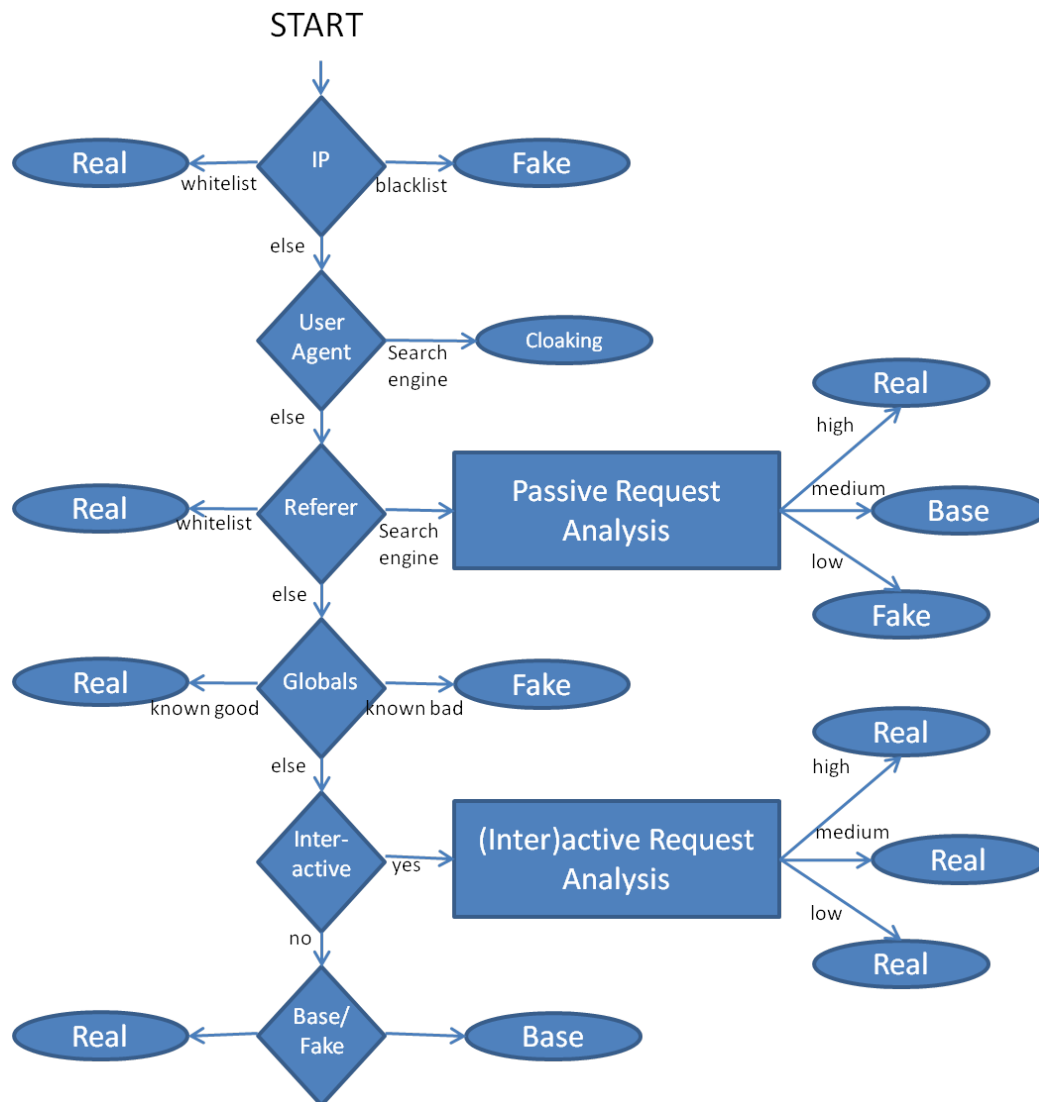


Figure 4.1: Partial Identity Composition Decision Tree.

referrer indicates that a search engine result was used to request the profile page the ranking algorithm will be used to determine which profile, i.e., how many real identity attributes, to present. This algorithm will be presented below in this section.

In the absence of a referrer **variables and their values within the URL** are considered for analysis. If available the values allow backtracking to the deployment by P , which can either be an invitation link or a hyper reference placed on another web site. In either case there are information stored within the back-end database that suggest either to present a fake profile or the real profile.

At this stage processing of all passively captured information is done and there are two options: Either the composition of a partial identity is finished or active request analysis is used to complement the intermediate result. In case of the first option depending on the configuration either a predefined basic profile or a fake profile is returned. The second option leads to processing of the results obtained via active request analysis, which in the same time would be the point of entry, whenever there is no passive request analysis employed.

The outcome of **active request analysis** again is a ranking based on the same algorithm (discussed below) as for the evaluation of search queries extracted from the referrer.

At the end of this phase (partial identity composition) in any case the outcome is a composition of a partial identity, which can be one of the following:

- Fake profile
- Basic profile
- Virtual partial identity
- Real partial identity

For the virtual partial identity the amount of real identity attributes and the overall quality of the profile depend on the ranking algorithm which is presented next.

Ranking Algorithm

As pointed out before the ranking algorithm is used for the evaluation of search terms used before being linked to the profile page of P or for the answers being provided during the (inter)active request analysis. The algorithm

Table 4.3: Identity attributes (ia) and their weights (w).

i	$ia_{i,j}$	j	w_i
1	first name	1,2	1
2	surname	1,2	1
3	city	2	2
4	place of work	1	1
5	job title	1	2
6	professional interests	1	3
7	private activities	2	3
8	associations	1,2	3

is a weighted ranking of identity attributes being provided by or known to the requester R . Each attribute, which matches a real identity attribute of the identity being requested, is weighted and then added to an overall sum. To distinguish between business and private related identity attributes, a vector is used. Thus, the final sum is a vector as well and is compared to configurable thresholds pairwise, which determine the amount of real identity attributes to provide (separately for business and private related identity attributes).

The ranking algorithm is a sum of weighted, matching identity attributes and can be written as

$$\begin{pmatrix} s_b \\ s_p \end{pmatrix} = \sum_{i=1}^n \begin{pmatrix} ia_{i,1} \\ ia_{i,2} \end{pmatrix} * w_i * p_i,$$

where $n = \text{amount of identity attributes}$ and $ia \in \{0, 1\}$: $ia_{i,1} = 1$, $ia_{i,2} = 0$ if ia_i is a business related *identity attribute*, $ia_{i,1} = 0$, $ia_{i,2} = 1$ if ia_i is a private related *identity attribute*, $w_i = \text{weight for the identity attribute}$ and $p_i = 1$ if there is a match for identity attribute i else $p_i = 0$, with $1 \leq i \leq n$. The resulting vector (s_b, s_p) contains the sum of weighted business (s_b) and private (s_p) identity attributes, which are provided during the initialization phase.

Since there is no experience with these measures yet, it is hard to define weights and thresholds in general. In the following we present numbers derived from our survey (compare Section 3.2) and verified by various test runs. Both, weights and thresholds depend on the configuration of the entire system and in particular on which identity attributes are in use. The provided numbers in Table 4.3 were successfully used in a minimal configuration.

Thresholds which lead to good results during our tests can be found in Table 4.4.

Table 4.4: Thresholds for sum of business and private related identity attributes (separately).

Threshold		Composed Partial Identity			
business	private	fake	basic	virtual	real
0	0	x	-	-	-
3	2	-	x	-	-
4	3	-	-	x	-
9	10	-	-	-	x

As mentioned before, the predefined basic partial identity (red) can be skipped. Instead thresholds for the virtual partial identity are extended accordingly. As a result we obtain two composed partial identities (one business- and one private-related) each of which is either fake, basic, virtual, or real. Merging these two partial identities will result in the final composed partial identity to be used for presentation to the requester. While the process for composing a fake, a basic or a real partial identity should be self explaining² the only question remaining is how the virtual partial identities are created with a relation of real and fake identity attributes corresponding to the outcome of the ranking algorithm.

Here we have to distinguish two different cases:

1. Given a basic partial identity and
2. absence of a basic partial identity.

In the first case we can determine the difference between the basic partial identity and the real partial identity for both, business- and private-related identity attributes. This range we call *virtual partial identity range*, i.e., a certain amount of real identity attributes complemented with fake identity attributes. Given the range for the outcome of the ranking algorithm for which we create a virtual identity (in Table 4.4 this range is 4 – 8 for business- and 3 – 9 for private-related ranking of a priori knowledge of R) this range can be mapped to the virtual partial identity range. Doing so for both business and private related identity attributes will allow us to compose a virtual partial identity with an amount of real identity attributes according to the outcome of the ranking algorithm. To allow also for reflection of different personal preferences we suggest to leave this composition measure configurable.

²Fake: composition of arbitrary fake identity attributes, basic: reuse of the (during initialization) predefined basic partial identity, real: composition of all real identity attributes provided during initialization.

In the second case, which is the absence of a basic partial identity, we can proceed the same way as in the first case, assuming the basic partial identity to be empty.

When it comes to selecting of real identity attributes among the same category, for instance in the case of more than one activity related to the identity under consideration, we suggest a random mode. Again personal preferences may become reflected via configuration.

4.5.5 Delivery of the Composed Partial Identity

The last phase is the easiest. Delivery of the composed partial identity happens by presenting the composed partial identity as a web site using a style template fitting the context, i.e., the environment where the partial identity is displayed. Whenever such a partial identity is returned to the requester R , the system will try to set a cookie to recognize R when requesting the same profile again in the future. Information on the composed partial identity will be stored in the back-end database for future lookup when the same cookie is used by R again. If setting a cookie fails, or the cookie is deleted between two subsequent requests, R might see different composed partial identities after each request. However, this will not bypass security of our proposed privacy measure as we will see later on in Section 4.7.

Issue with Fake Identity Attributes

Since fake identity attributes might be misleading or even harmful to the reputation, we also foresee a configuration which solidly varies the amount of real identity attributes without complementing them to a composed partial identity using virtual identity attributes.

4.5.6 Results

The implementation presented here instantiates our dilution approach using two different dimensions: Business and privacy. During our research we explored different strategies here using only one or even more than two dimensions, where each strategy comes at its advantage. In this work we decided for two dimensions in order to keep the underlying model as simple as possible, but not hiding the overall potential. In the remainder of this subsection we discuss our achievements to display the advantages of our model and justify our design decisions.

The terms privacy and private appear to be very related at a first glance. In depth reasoning leads to the conclusion that privacy is a right being rele-

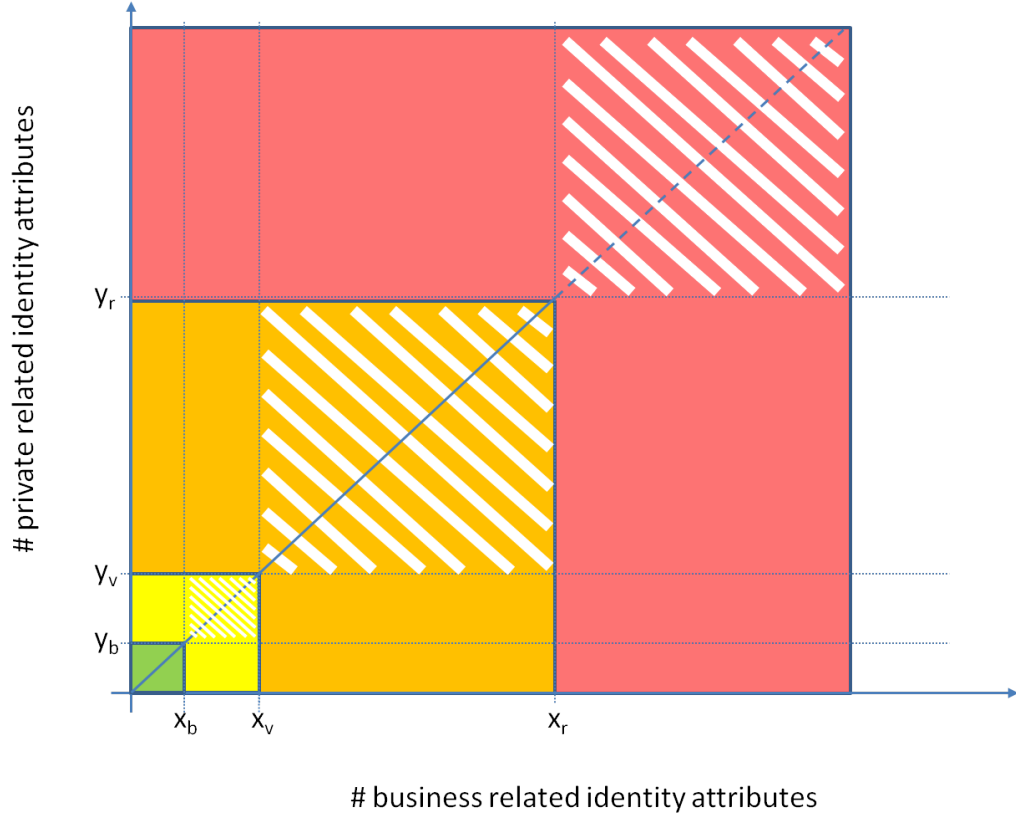


Figure 4.2: Two dimensional privacy graph.

vant over multiple domains, which private life is one of. The most dominant domains are business and private life. That is why we choose them in our implementation of dilution. Accordingly we came up with a two dimensional privacy vector field. Further domains, which can be distinguished include: general information, sports, holidays, family, friends, digital pictures. Thus, inducing further domains may result in a partitioning of private life (the domain used in our implementation presented above), but may also add entirely new domains. The decision here is heavily depending on the comprehensiveness of the real partial identity and the presentation of it. Additionally, it also relates to the environment, where it will be used, e.g., a social online platform, a personal web site, or a business web site.

The two dimensional approach implemented above can be best understood by looking at the graph shown in Figure 4.2.

On the x-axis we have a business, on the y-axis a private related measure as given by the ranking algorithm. A request can be either balanced between the two dimensions, i.e., being mapped to the diagonal line, or tend to either

be more business or private related. Gradient of the diagonal depends on the amount of available information and particularly on the relation between the different dimensions (private and business). It is not an absolute border but rather a loose bound between the two dimensions/trends. Assuming there are more business than private related identity attributes provided by the identity owner / publisher P , then the gradient is accordingly smaller. The gradient g can be calculated as follows:

$$g = \frac{\# \text{ private related identity attributes}}{\# \text{ business related identity attributes}}$$

Correspondingly, a certain real partial identity with for instance ten business related identity attributes and five private related identity attributes will result in a composed partial identity with a trend to an overlap in real private related identity attributes already for a measure of (3,5). Note, that thresholds can but do not have to correlate to the gradient.

In Figure 4.2, we assume a correlation between thresholds and the gradient of the diagonal telling private and business related domains apart. Thus, these thresholds are represented by points on the diagonal. Therefore, x_b marks the threshold for the optional basic business partial identity, x_v for the virtual business composed partial identity, and x_r for the real business partial identity. Respectively y_b , y_v , and y_r represent the thresholds for the private partial identities.

4.6 Application

The proposed concept has been applied in two different scenarios – an online social network and a personal homepage. While in the first one the interactive request analysis is realized, the latter one purely relies on passive request analysis. Both applications work as expected and will be presented in the following.

4.6.1 Online Social Networks

The interactive approach has been build on top of PHP-Fusion [47] an open source content management system providing different social network features like

- Membership management
- Personal profile pages

- Message exchange
- Creating and joining customized user groups
- Picture upload
- Search functionality

In the used version of PHP-Fusion (version 7), the search algorithm is very basic. A given input to the search algorithm is used as one string and gets compared to all user names. If one or more matching users are found, links to their corresponding profile pages are returned.

In order to integrate the dilution methodology the search has been modified in a way that different identity attributes can be entered in the search form separated by a “+” character. Hence the database lookup is extended accordingly. All datasets are compared to each identity attribute entered in the search form. The resulting list of user profiles with a match is then further processed. Depending on the categories of matching identity attributes, a decision will be made if the requester has rather a business or private relation with the identity owner being searched for. The amount of matching identity attributes is compared to a configurable threshold each user of the social network can decide individually. A third aspect being considered is if there is a group which both requester R and identity owner/publisher P belong to. If this is the case a relation of real and virtual identity attributes, which can be configured by each P separately, is used to decide on the identity attributes used during the partial identity composition phase. If P and R do not have a group membership in common the matching categories (here private and business) are evaluated to decide whether R has a priori knowledge about P rather in a business or in a private context. Here also the amount of matching attributes is taken into account. The links being returned in response to the search query point to a landing page which will use questioning for further business or private related identity attributes. The questions depend on whether business or private related knowledge about P has to be confirmed. The answers provided by R will be used as input for a second run of the ranking algorithm. Finally, enough certainty on the individual being searched for is given so the partial identity can be composed and presented as a user profile web page. The overall process is shown in Figure 4.3.

Implementation Details

To achieve the above sketched behavior, several changes to PHP-Fusion are necessary and additional processing is required.

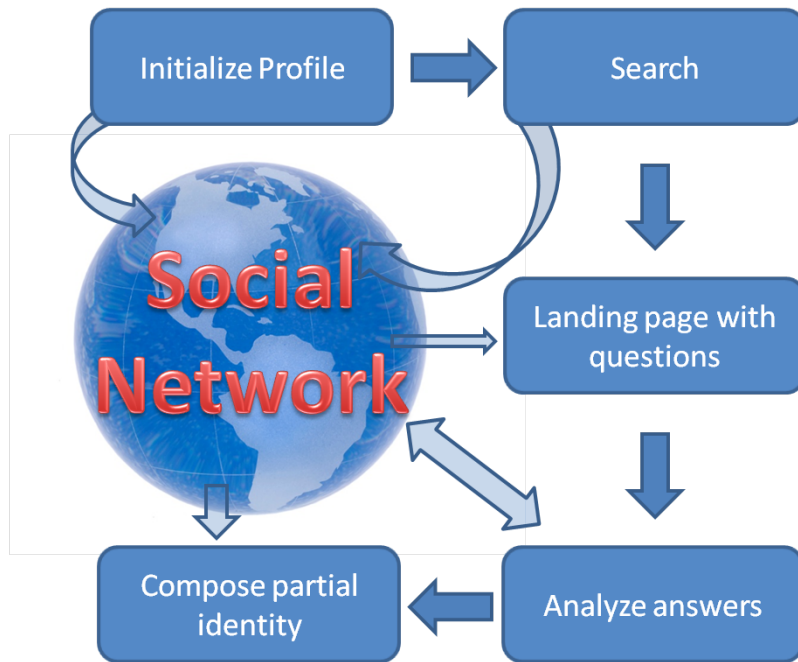


Figure 4.3: Overview interactive dilution process applied in an online social network scenario.

The existing table *fusion_users* is extended to become capable of holding more personal information, i.e., identity attributes.

An additional table *fusion_criteria_ranking* (compare Table 4.5) is added where each of the identity attributes is categorized to be either personal, general, or contact information. Furthermore a ranking value is added to each identity attribute. This value can be defined by the user during account/profile creation via a dedicated input form. Last a question string is added to each identity attribute asking for the corresponding information, e.g., what is the real phone number of this user.

Another table *fusion_validity_values* is added to hold validity information for a virtual partial identity given in percent – see Table 4.6.

To enable predefined virtual partial identities another table is created (*fusion_fake_profile*). This table can be filled by the user via a dedicated form during account creation. Once the virtual identity attributes for a virtual partial identity are submitted the system will calculate the corresponding validity in percentage: Comparing the virtual partial identity and the real partial identity attribute-wise, the amount of matching, i.e., real identity attributes within the virtual partial identity is divided by the sum of available identity attributes. Additional information being stored here are the amount

Table 4.5: Categorization and ranking of different identity attributes. PI: personal information, GI: general information, CI: contact information, UR: undefined ranking.

id	ranking_name	value	ranking_description	additional_question
0	undefined_rank	0	UR	undefined
1	user_name	1	PI	user name of this user?
2	user_real_name	1	PI	real name of this user?
3	user_tel	1	PI	real phone number of this user?
4	user_location	2	GI	where is the user from?
5	user_dateofbirth	1	GI	when was the user born?
6	user_job	2	GI	job of this user?
7	user_sig	2	GI	signature of this user?
8	user_email	3	CI	email address of this user?
9	user_aim	3	CI	aim id of this user?
10	user_icq	3	CI	icq number of this user?
11	user_msn	3	CI	msn-nick of this user?
12	user_yahoo	3	CI	yahoo nick of this user?
13	user_web	3	CI	web site of this user?
14	user_workplace	2	GI	workplace of the user?
15	user_hobby	2	GI	hobby of the user?

Table 4.6: Discrete separation of different validity levels.

validity_id	validity_in_percent	validity_description
1	0	0% valid information
2	20	1-20% valid information
3	40	21-40% valid information
4	60	41-60% valid information
5	80	61-80% valid information
6	100	81-100% valid information

Table 4.7: Database table to store fake profiles along with corresponding validity information.

profile_id	user_id	fake_name	fake_mail	fake_tel	...	fake_validity	fake_validity_value	fake_avatar_validity
1	10	John Doe	John.Doe@sample.org	012345678	...	20	3	0
2	11	Pseudo Nym	secret@imaginary.org	666-666	...	60	8	0

Table 4.8: Extension of *fusion_user_groups* table by an *group validity* field.

group_id	group_name	group_description	group_validity
5	university	fellow students	60
6	friends	close friends	80

of real identity attributes and the validity of an provided avatar. The latter one can be 0 or 1, which means *false* or *true* respectively, and has to be provided by the user during virtual partial identity profile submission. The resulting table is given in Table 4.7.

Following our goal to reflect real-world relationships and privacy decisions, different trust levels can be configured for each user group, which is supported by an added *group_validity* field (see Table 4.8). For instance a new group *closest friends* might get 100 percent assigned as a group validity, wherein another group *colleagues* might get only 60 percent group validity.

Based on these modifications and additional information all search requests are stored in a new table named *fusion_calculate_profile_validity*. Table 4.9 shows the corresponding data representation within the database. In this table not only the amount of matching identity attributes between the identity attributes provided during the search and a corresponding user's account real identity attributes (*hit*) are stored, but also the exact identity attributes which actually match (*hit_string*). Further information is automatically calculated by looking up the rankings (given in Table 4.5) of matching identity attribute and filling this into the *hit_ranking* field, and calculating the percentage of real identity attributes provided during the search

Table 4.9: Database table to calculate validity of a profile being returned in response to a given query.

calc_id	user_id	imemeber_id	hit	hit_string	hit_ranking	calc_percentage	calc_validity	search_query	additional_information_hit
1	11	10	1	0.0.0.1.0.0.0.0.0.0.0.0.0.0.0	0.2	13.33	50	John+USA	5
2	10	10	2	1.0.0.1.0.0.0.0.0.0.0.0.0.0.0	1.0.2	13.33	100	John+USA	3
3	12	10	1	1.0.0.0.0.0.0.0.0.0.0.0.0.0.0	1.0	6.25	100	Joejoe	0
4	1	10	0	0.0.0.0.0.0.0.0.0.0.0.0.0.0.0	0	25.00	0	John+USA +1970-01-01	0

calc_percentage. For determining the calculated validity (*calc_validity*) first R and P are tested for having a group membership in common. If this is true, the corresponding group validity, as defined by the group founder (compare Table 4.8), is used. If R and P have more than one group membership in common, the group with the highest group validity is considered. If there is no common group, then the validity is calculated via amount of real identity attributes provided within a query divided by the amount of strings being used in the query.

Given all the above discussed information the amount of rankings can be counted. The ranking, which appears most often, is used to assemble the before mentioned landing page asking further information on identity attributes according to the most prominent ranking. Depending on the configuration of the system more than one question can be asked here. All questions, which are answered right, result in a value entered in the *additional_information_hit* field of the *fusion_calculate_profile_validity* table. The way answers can be provided again depends on the configuration. Answers can either be chosen from a drop down menu or entered in a free text form.

Finally, all the gathered or calculated information can be used to present a profile page with a corresponding validity, based on the partial identity composition, which depends on thresholds to be defined before. The identity attributes used for the partial identity composition are either real or virtual. As virtual identity attributes for a certain individual P , real identity attributes of other registered users are taken if P has not predefined virtual partial identities, which should be used instead. As a result virtual

partial identities can contain real partial identity attributes of different users. The advantage here is that we do not necessarily need to provide additional identity attributes.

Obviously, the amount of registered users is crucial for the system to work. If there are only very few registered users, there are too less identity attributes for a appropriate dilution. In this case we depend on additional identity attributes to be provided or enough predefined virtual partial identities to be available. To guarantee this we suggest to have a minimal amount being defined, which determines the amount of virtual partial identities to be defined during initialization of the system in order to successfully register. Of course, this lowers usability a lot. However, this is not considered as a drawback if the system works once enough users are registered, since most online social networks have millions of users anyway. Robustness of this implementation under the assumption of many registered users will be discussed in Chapter 5. It should be mentioned that dilution of any given identity using real identity attributes of other identities, does not prevent harvesting real identity attributes of course. Hereby, only linkage/aggregation of different identity attributes to a real identity is prevented.

4.6.2 Personal Homepages

In an online social network context user interaction, i.e., user-to-user and user-to-platform, is a key element and hence we expect users to accept additional interaction as it is caused by the active request analysis employed in our application example presented in Section 4.6.1. When it comes to personal homepages the scene changes. Users want to browse information without being asked for additional information as in the previously presented application. Of course, acceptance of interaction with personal homepages might increase if it becomes more established. For the moment we present an example application solidly relying on passive request analysis, which is presented in the remainder of this section.

To prove functionality of passive request analysis in the context of dilution, we have implemented a profile generator which can be deployed by any user in order to dilute personal information presented on a personal web site, e.g., a profile page. While profile pages within online social network platforms are less (or not at all) visible to search engines, the opposite is the case for personal homepages – assuming no additional protection measures like access control are in place. Even though it is hard to find precise numbers here, there are statements claiming that *“70 percent of the traffic to most web sites is referred by Google”* (Source: http://searchengineoptimism.com/Google_refers_70_percent.html). Therefore, it appears to be reasonable

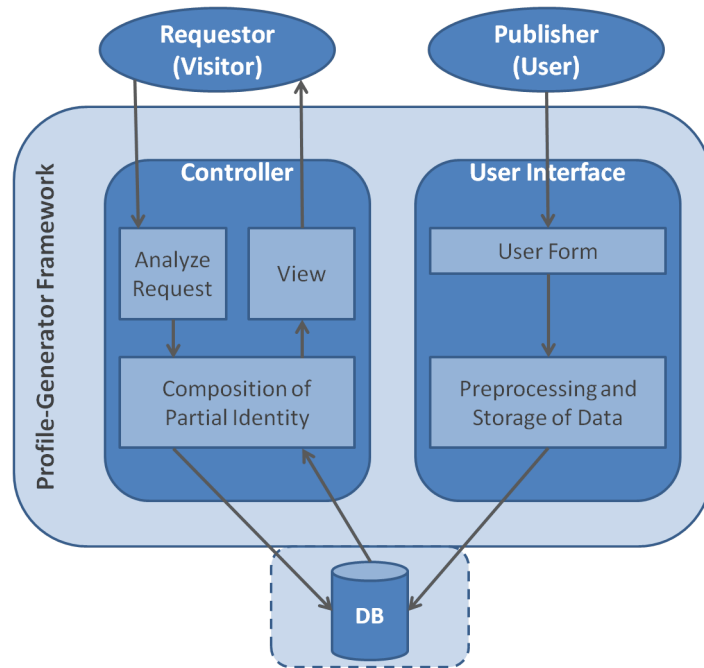


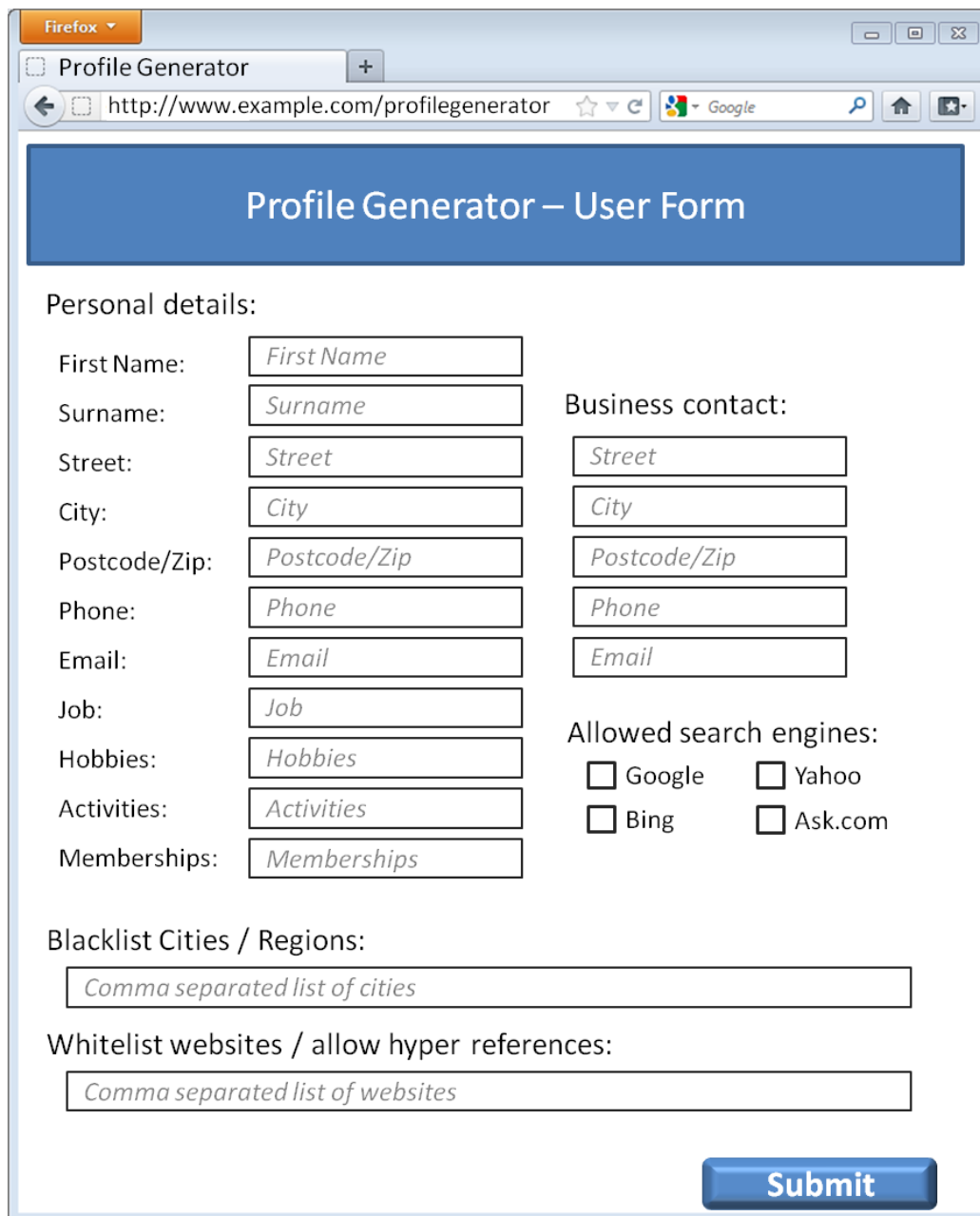
Figure 4.4: Overview Profile Generator.

to apply our passive request analysis based approach with a strong focus on analyzing the **Referer** header.

Figure 4.4 displays an overview of the system presented in this section. A user P employing the profile generator initially fills a form with his real identity attributes. The provided data is processed and then stored in a database. Whenever a requester R is trying to access the personal homepage a controller will start passively analyzing the request (*passive request analysis*, compare Section 4.5.3). Next, a partial identity is composed based on the outcome of the passive request analysis and the available data within the database. The resulting partial identity is then displayed to R .

Implementation Details

To run this application a PHP capable web server with a database is required. Once the application is installed the user, who wants to use the profile generator in order to dilute its personal information, is asked to provide its real identity attributes via a form shown in Figure 4.5. Beside private and business related real identity attributes the user can choose, which search engines should index the personal homepage, to which cities (geoIP location of requester) no real identity attributes will be displayed, and which web sites are



The screenshot shows a web browser window with the title "Profile Generator" and the address bar displaying "http://www.example.com/profilegenerator". The page has a blue header with the text "Profile Generator – User Form". Below the header, the form is organized into several sections:

- Personal details:** This section contains a vertical list of input fields on the left: "First Name:", "Surname:", "Street:", "City:", "Postcode/Zip:", "Phone:", "Email:", "Job:", "Hobbies:", "Activities:", and "Memberships:". Each field has a placeholder text corresponding to its label. To the right of these fields is the "Business contact:" section, which also contains a vertical list of input fields: "Street:", "City:", "Postcode/Zip:", "Phone:", and "Email:", each with a placeholder text.
- Allowed search engines:** This section is located below the business contact fields and contains four checkboxes with labels: "Google", "Yahoo", "Bing", and "Ask.com".
- Blacklist Cities / Regions:** This section is located below the search engines and contains a single text input field with the placeholder text "Comma separated list of cities".
- Whitelist websites / allow hyper references:** This section is located below the blacklist and contains a single text input field with the placeholder text "Comma separated list of websites".

At the bottom right of the form is a blue "Submit" button.

Figure 4.5: Profile Generator user input form.

allowed to link to the real partial identity.

All submitted identity attributes are extended with synonyms before being stored to the database. This improves matches in case of imprecise or related query terms being used by requesters searching for the profile of a

Table 4.10: Simple weighting for personal homepage profile generator.

Identity Attribute Category	Weight (Factor)
Name	1
City	2
Work place	1
Job title	2
Interests	3
Activities	3
Membership in associations	3

Table 4.11: Thresholds for partial identity composition.

Composed Partial Identity	Threshold
all real identity attributes	6
business related information only	3

certain individual. Finding synonyms is done in an automated fashion using public web services like Thesaurus [21]. Here also other web services like dictionaries, etc. could be employed in the future.

Since the implementation of this application is very close to the concept presented in Section 4.5, we only highlight certain details and differences comparing to the general concept.

In this application we do not make use of (inter)active request analysis. Apart from this difference the decision tree is as given in Figure 4.1. The strongest measure in this approach are search terms extracted from the **Referer** header in the likely case that our profile has been linked within the results returned by a search engine. Then the URL in the **Referer** header contains all search words as shown in Table 4.1. The ranking algorithm in use is very basic: Each identity attribute category is assigned with a weight and for each match these weights are summed up. The corresponding weights are presented in Table 4.10.

Also for the thresholds we use a very basic setting. The corresponding values being used for the partial identity composition are displayed in Table 4.11.

Of course, we may have implemented a more sophisticated ranking (compare Section 4.5.4). Nevertheless, in this application we would like to point out the robustness of our proposed dilution approach already for very basic correlation. We will discuss functionality of this approach in Section 4.7. Further evaluation will be presented in Chapter 5, where we illustrate the robustness of this approach facing different attack vectors.

4.7 Functional Evaluation

In this section we focus on the functionality provided by the dilution methodology. Therefore, we favor to point out the capabilities of our approach along with practical applications and postpone robustness considerations to Chapter 5.

The comprehensiveness of our approach is described in Section 4.5. We present a concept including different means to preserve privacy following a methodology we call dilution. The key idea is to transfer privacy means from the physical into the digital world. This particularly involves presentation or sharing of different partial identities depending on the given context or environment. Our approach enables adjustable granularity to a large extent. It cannot only distinguish between different categories of identity attributes, but also allows for multi-dimensional privacy measures. While describing the concept we stick to two dimensions, i.e., private and business related, for the sake of simplicity and in order to picture the resulting privacy control in two dimensional graphs. However, the method is explained in all details necessary to transfer the approach into three or more dimensional space whenever required. The employed techniques for information extraction can be scaled from very convenient usability (*passive request analysis*) to robust reliability (*active request analysis*). The design enables sophisticated ranking algorithms, but has shown to serve sufficient protection in very basic implementations as given in two application scenarios. Deployment of the proposed methodology is not necessarily depending on major restructuring of online social network providers, but can happen on a single user base, e.g., for personal homepages, as well.

While presenting a comprehensive solution for a very prominent and current problem, we also illustrate the most simple and straight forward application of it. Rendering all variable or configurable components of our concept to a minimum in terms of complexity, we still face a sufficient working solution.

Before turning to the actual evaluation we will discuss a reasonable deployment of both dilution applications, i.e., in a online social network and as a profile generator for profile pages as a part of general personal web sites. Commonly, online social networks expose most of the contained information to other members only. If this does not hold then the online social network application is just a special instantiation of the personal homepage application. Thus, we decided to test the online social network application within a closed network environment wherein we simulated different user interactions and the personal homepage application on a live system, which is connected to and accessible from the Internet.

The first objective we evaluate is the privacy threat posed by **harvesters**,

crawlers, spiders, or similar technologies. Their only purpose is automated exploration of web sites in the Internet with the intention to identify, steal, and collect valid identity attributes. In the worst case this can even lead to identity theft (discussed in Section 2.5). In all other cases at least abuse of these information is likely to happen and results for instance in unsolicited bulk emails. The personal homepage application of dilution as presented in Section 4.6.2 was extensively tested to be resistant towards such threats. This test was conducted as described in the following. During the initialization process where the user has to fill in a form with real identity attributes two different email addresses were submitted, one as private and one as business related. While both of these email addresses were only shown in response to corresponding requests a third email address was shown unconditionally at any request. For all three email addresses, real accounts have been created. Throughout the testing period of several weeks no single unsolicited email has been received on neither the private nor the business related address. The third email address, which was unconditionally shown on any request received 4 unsolicited emails. Even though this is not a very strong result, given the uncontrollable variables, e.g., an email harvester might crawl our web site but there is no guarantee, this finding confirms our expectation and serves as an indicator for success of our approach, in terms of protection against harvesters, crawlers, and spiders.

Use of **cloaking** – presenting different content to search engine spiders than to regular users – also worked as expected. All search engine bots were successfully identified and got a special representation of identity attributes, namely showing all identity attributes in a mixed fashion. As a result all identity attributes (real and virtual) will be indexed by the search engines. This was verified by waiting until a search bot has accessed our web site and then checking the search engine’s cache. A screen shot of Google’s cache of a web site generated with our personal homepage application can be seen in Figure 4.6.

Here, we see different identity attributes being arbitrarily assembled to user profiles which are then listed. Real identity attributes are highlighted. Although this view does not help a requester to figure out which identity attributes are real and which are not-real, we decided to turn off caching by adding a corresponding meta tag to our web site, which is respected by search engines. The short description below each link returned by a search engine is hereby not affected. Actually, the property of the short description, being assembled of matching lines of the target web sites, is even supporting our approach.

Also the **active request analysis** as implemented in the online social network application works as expected. Queries are handled according to

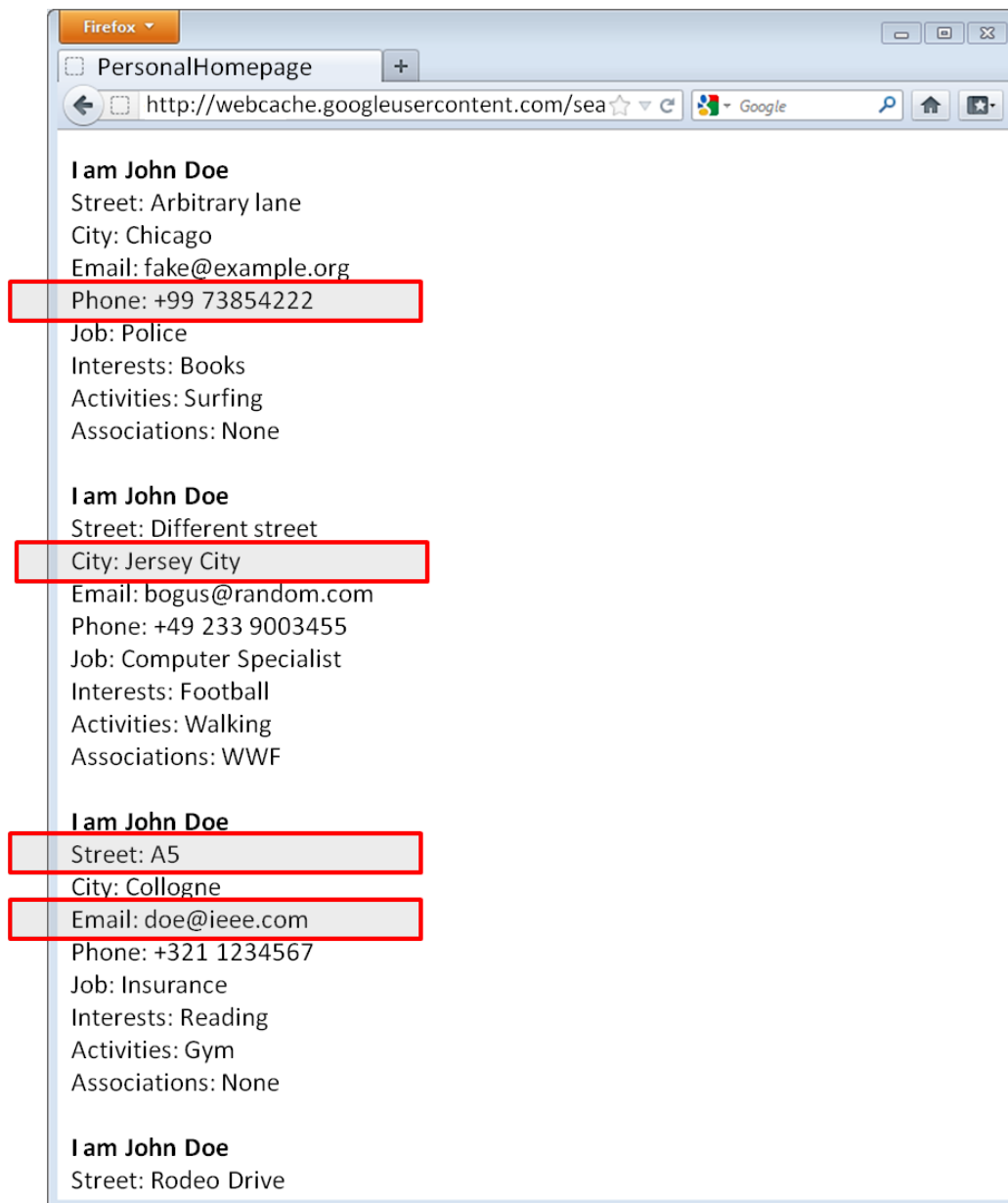


Figure 4.6: Snippet of Google's cache content. Real identity attributes are highlighted.

the defined thresholds. Knowing these we were able to successfully submit queries with the intention to receive a particular response. Responses could be triggered to return real, fake or basic partial identities. Also (inter)active questioning was triggered successfully and turned out to take impact on vague ranking algorithm results in both directions, return of real or fake partial identity, as configured before.

4.7.1 Limitations

While almost all designed and implemented functions were evaluated as expected, we also identified two limitations we do not want to withhold. Both applications have shown that the entire approach heavily depends on the **availability of a massive pool of virtual identity attributes** for all identity attribute categories. In case of virtual partial identities being composed using real identity attributes of other users, the overall amount of participating users is the critical parameter for dilution to work for instance within a online social network application. If there is only one user trying to dilute its personal information a sufficient pool of virtual identity attributes is required. Nevertheless, this limitation can be overcome by either enough participating users or a general public available pool of virtual identity attributes, which can be reused by different installations independent from the environment the system is deployed to.

The other weak point we identified is the use of **geoIP lookup services** in order to deny access to the real personal identity attributes of a certain individual for all request from a certain city. Generally geoIP lookup services are not precise enough to provide an accurate measure for this purpose, because they do not locate the actual end-user but rather the providers up-link to the Internet.

Apart from these to minor limitations the system was found to work as expected and can conveniently be used in the scenarios given for our applications.

4.8 Conclusion

In this chapter we have extended our set of definitions by adding *mimetic* and *polymorphic dilution*. Next we identified phenomena in the past closest to what we refer to as dilution. After sketching our idea of dilution we presented a full design concept and relevant implementation pillars. Here we particularly elaborated on active and passive request analysis, and the partial identity composition for returning a response corresponding to the

request. Last we present two sample applications (personal homepage for passive request analysis and social online network for active request analysis) which we fully implemented to proof our concept and evaluate all functions. The functional evaluation of the implementations has been presented in this chapter as well.

Chapter 5

EVALUATION OF DILUTION

5.1 Introduction

In Section 4.7 we already evaluated functionality of our concept as implemented in two different applications. We have seen that the system works as expected. In this chapter we further stress test our applications to evaluate whether the approach is robust enough so it cannot be bypassed by an attacker easily. The threat model is simple: An attacker with limited knowledge on the personal information of a certain individual under consideration tries to enumerate all its available, real identity attributes using an unlimited amount of arbitrary requests. The attacker is successful if she manages to learn about a new identity attribute given a certain set of real identity attributes such as the likelihood of the new identity attribute being a real identity attribute of the individual under question, is probabilistically significant. We do not consider here finding additional real identity attributes from other web sites than the one implementing dilution, because this would cause the system to disclose further information as it is meant to do by design. Since we cannot execute an unlimited amount of request, we decided to have 2000 requests for each strategy. After several empirical tests, this appears to be sufficient enough to identify whether real identity attributes have a higher frequency than virtual/fake ones. Before turning to the actual attack evaluation we would like to mention that all identity attributes presented in this section were mapped from the original set to an imaginary set to assure that no personal information can be disclosed by being presented in this thesis. The mapping has been performed in a way such that no results will be altered by using bijective projection.

Table 5.1: Imaginary identity attributes used as real partial identity.

Category	Identity Attribute	Weighting
Name	John Doe	1
City	Jersey City	2
Work place	New York	1
Job title	Computer Security Specialist	2
Interests	Programming, Firewalls	3
Activities	Sailing, Reading, Squash	3
Membership in associations	IEEE, ACM, FIRST	3

5.2 Attacking Passive Polymorphic Dilution

5.2.1 Strategy

First, we investigated the evaluation of our passive polymorphic dilution application presented in Section 4.6.2. Therefore, we created an imaginary partial identity serving as a real partial identity as shown in Table 5.1. For the sake of completeness weighting is also listed as chosen in Section 4.6.2.

Since cookies are used to present the same partial identity to subsequent requests by the same requester, we disable cookies. Furthermore, we do not exclude any geographic locations using geoIP lookups, assuming the attacker is requesting our web site from an allowed location. Also invitation links are not used throughout the testing, since those are meant to disclose the real partial identity as provided during initialization of the system. As a user agent we employ varying legitimate *User-Agent* strings not being used by search engines. Variables within the URL as they might be used in invitation links are not part of the evaluation either. The intention of our brute-force approach aiming to reveal the real partial identity is thus to produce combinations of different *Referer* header values, wherein we combine variables as used in search-engines. This way we emulate different requests preceded by search engine queries using different identity attributes as search terms. Conducting the automated brute-force attacks we use five different strategies, each of which performing 2000 subsequent requests with varying search terms transfered in the *Referer* header field. The different strategies are described in the following.

Strategy 1: The first request sent to the web service hosting our personal homepage dilution application uses a referrer indicating a preceding search query for the first name of the individual presented in the corresponding personal homepage. The resulting profile is parsed, all identity attributes

shown are stored in a database and are compared to the real partial identity. For the next round, i.e., the next request, a random selection of the before parsed identity attributes in combination with the first name is assembled to a referrer as it would be transmitted when using the same identity attributes in a search engine query. This algorithm is repeated 2000 times or until the returned profile page is equal to the real identity used to initialize the system. The latter condition raises the bar of robustness tremendously, since a single random hit would prove our approach to fail.

Strategy 2: This strategy is similar to Strategy 1, with the difference that the set of identity attributes randomly chosen from in order to assemble the referrer string is limited to the most frequently seen, i.e., the most frequently returned identity attributes, during the previous rounds/requests.

Strategy 3: Like in Strategy 2, the frequency of identity attributes being returned is considered, but this time also the weighting of the identity attribute categories is known (compare Table 4.10). Identity attributes of higher weighted categories are preferred during the composition of the referrer.

Strategy 4: Here, the previous strategies are combined. The first 500 iterations are conducted as in Strategy 1. For the remaining 1500 iterations procedure is as in Strategy 3.

Strategy 5: Last, we conduct the attack like in Strategy 4, but this time combining the random selection of identity attributes not only with the first name but additionally also surname and profession are submitted in each request within the referrer.

5.2.2 Results

By running the brute-force attack as described in the previous section 2000 profiles, i.e., virtual partial identities, were harvested by each strategy. In Table 5.2 we present the 20 most frequent hits (identity attributes) for each strategy, along with the amount of corresponding hits and the weight of the category if considered.

Table 5.2: Results of brute-force attack with five different strategies. Real identity attributes marked in green.

Strategy	20 most frequent identity attributes	# hits	Weight
Strategy 1	John	1884	
	Doe	1884	
	IEEE	734	
	VVPP2009@yahoo.de	435	
	inf907@gmail.com	426	
	345name@gmail.com	417	
	1234254@hotmail.de	408	
	sk8976@web.de	403	
	m19@yahoo.com	393	
	email2345@yahoo.com	393	
	cone345@hotmail.com	389	
	reading	376	
	sun3425@web.com	375	
	k1990@hotmail.com	363	
	Berlin	360	
	Berlin	337	
	singing	233	
	Washington	232	
	Washington	219	
	8909765	216	
Strategy 2	John	1879	
	Doe	1879	
	cone345@hotmail.com	434	
	1234254@hotmail.de	426	
	VVPP2009@yahoo.de	404	
	inf907@gmail.com	403	
	email2345@yahoo.com	400	
	m19@yahoo.com	394	
	sun3425@web.com	391	
	345name@gmail.com	387	
	sk8976@web.de	384	
	k1990@hotmail.com	379	
	poetry	204	
	3244444	203	
	reading	201	
continue next page			

<i>Continuation of previous page</i>			
Strategy	20 most frequent identity attributes	# hits	Weight
	567483	199	
	4563728	198	
	345267	198	
	writing	197	
	listing music	196	
Strategy 3	John	1885	1
	Doe	1885	1
	IEEE	1604	3
	reading	1327	3
	Berlin	979	1
	Berlin	917	2
	Transportation Engineer	797	2
	Interpreter	471	2
	345name@gmail.com	445	1
	sun3425@web.com	435	1
	email2345@yahoo.com	410	1
	1234254@hotmail.de	406	1
	inf907@gmail.com	405	1
	VVPP2009@yahoo.de	402	1
	sk8976@web.de	382	1
	cone345@hotmail.com	378	1
	k1990@hotmail.com	374	1
	m19@yahoo.com	365	1
	Chicago	299	1
	Chicago	279	2
Strategy 4	John	1876	1
	Doe	1876	1
	IEEE	1367	3
	reading	1076	3
	Berlin	850	1
	Berlin	795	2
	Transportation Engineer	754	2
	email2345@yahoo.com	436	1
	inf907@gmail.com	418	1
	m19@yahoo.com	414	1
	cone345@hotmail.com	406	1
	k1990@hotmail.com	405	1
<i>continue next page</i>			

<i>Continuation of previous page</i>			
Strategy	20 most frequent identity attributes	# hits	Weight
	sk8976@web.de	390	1
	345name@gmail.com	385	1
	sun3425@web.com	384	1
	VVPP2009@yahoo.de	383	1
	1234254@hotmail.de	381	1
	Chicago	290	1
	dentist	286	2
	Chicago	274	2
Strategy 5	New York	1966	1
	John	1886	1
	Doe	1886	1
	Computer Security Specialist	1884	2
	New York	1312	2
	IEEE	1000	3
	reading	470	3
	345name@gmail.com	427	1
	cone345@hotmail.com	420	1
	inf907@gmail.com	419	1
	sun3425@web.com	413	1
	email2345@yahoo.com	410	1
	k1990@hotmail.com	392	1
	VVPP2009@yahoo.de	391	1
	1234254@hotmail.de	384	1
	sk8976@web.de	373	1
	m19@yahoo.com	373	1
	singing	223	3
	Berlin	219	2
	45362353	206	2

As we can see, only Strategy 5 was successful in disclosing further personal information. In all other strategies the available information on the individual being searched for could not be extended significantly. The only certainty that can be derived by frequency analysis is the name and a membership in one association. In Strategy 2 we could not even identify the association membership. In Strategy 5 revealing more personal information does not point to a weakness of our approach, but behaves aligned to the design idea: The returned results should be corresponding to the a priori knowledge of the requester. Here, we impersonated a requester with certain

knowledge about the profession of the individual under consideration. As a response business related real identity attributes have been returned more frequently. Thus, not only the job title, which was used within the query, but also the work place (here: New York) has been returned. Noteworthy is that the city where John Doe is living has not been exposed. Instead the city has been diluted with the work place, namely New York. The two different categories of identity attributes can be told apart by comparing the results with the weights as defined in Table 5.1. Therefore, the most frequent hit in Strategy 5 (New York) is a real business related identity attribute, while fifth most frequent identity attribute is a virtual/fake identity attribute for the residence of John Doe. Actually, the real identity attribute for the residence (New Jersey) is not present among the 20 most frequent hits. Instead we see another prominent dilution for this category being Berlin with 219 hits.

5.3 Attacking Active Polymorphic Dilution

5.3.1 Strategy

In order to address the different approach in implementing active polymorphic dilution as described in Section 4.6.1, we change our strategy appropriately. The given implementation of a social network dilution approach favors reuse of real identity attributes belonging to other members of the same on-line social network platform over inducing additional fake information. Still, virtual identity attributes can be entered into the system by creating explicit virtual partial identities, which will be used during the partial identity composition phase. Since each search for a particular person will lead to a landing page asking for further identity attributes, which then can be chosen from drop-down menus, the strategy in use is as follows.

By subsequent requests all identity attributes that can be chosen from the drop down menus can be harvested. Therefore, the amount of requests here, depends on the available identity attributes. For the sake of simplicity and completeness, we are assuming the worst case where an attacker is able to grasp all available identity attributes. With this information at hand, we can automate downloading composed partial identities for all combinations of answers provided on the landing page. The profile we are searching for belongs to John (compare Table 5.1) again. The only difference here is that John also has a user name within the social network, namely JohnDoe. So this is the nickname we provide as a search term. To assure that we do not monitor any special case, e.g., due a very common name shared by different members, we made sure that there is no other user with similar or same

Table 5.3: Top 20 after frequency analysis of brute-forcing results.

20 most frequent identity attributes	# hits	Weight
horse riding	11340	1
anna	7710	1
Smith	7614	3
Germany	7610	1
Spain	7567	1
herderl_derdel	7561	3
www.example.com	7545	2
hurrdurr	7521	1
145987654	7494	1
fastu@xicht.de	7483	2
fastu	7483	3
123214	3949	1
Sammy	3939	1
BMW	3921	3
frankie44827@yahoo.de	3920	2
www.peter-online.de	3878	2
herp@derp2.com	3869	2
emily_msn	3866	3
peter@t-online.de	3864	2
Traveling	3856	1

name.

5.3.2 Results

Following the strategy as described in the previous section we obtain the top 20 hits presented in Table 5.3.

As we can see there is not a single match between the top 20 results (according to frequency) and our imaginary set of real identity attributes as listed in Table 5.1. Trying different variations of the strategy we were not able to improve the results presented here. Therefore, we desist from presenting further strategies here, but instead rather discuss some general issues we realized during our testing.

The implementation of polymorphic dilution within an online social network platform as presented in Section 4.6.1 appears to be robust, i.e., it is impossible to derive the real partial identity of the considered user by a brute-force, frequency attack as we applied it. Still, a harvester could easily

grasp loads of real identity attributes, which are presented as choices within the drop-down menus. As a result email addresses could be extracted and used for malicious scam campaigns. Although identity theft is countered, the leakage of valid email addresses is of major concern.

During our tests we figured out using real identity attributes of other users to dilute the profile of a certain user is not the only problem here. There are further drawbacks in using drop-down menus, for evaluation of a priori knowledge of the requester. Instead of using real identity attributes of other registered users, we might simply use virtual identity attributes surrounding one real identity attribute. Choosing the latter one a requester will increase its ranking which impacts on the composed partial identity being presented. Choosing one of the virtual identity attributes will result in a lower ranking. The problem here is what virtual identity attributes to present. Using a certain fixed set will allow an attacker to search for different users and compare the choices within the drop-down menus displayed on the landing page. Those identity attributes which differ can then be identified as real identity attributes of the corresponding profile requested. Configuration of the algorithm to use different sets of virtual identity attributes to dilute the real identity attribute also displayed in the drop-down menu will allow an attacker to perform subsequent requests for the same user looking for the identity attributes to stay identical and hereby determining the real identity attributes of the user under request. Both ways, iterative queries will disclose the real partial identity of the requested user's profile. For this reason an implementation using drop-down menus requires more sophisticated strategies to not have an attacker easily bypassing the system.

Alternatively, answers could be provide by plain text input forms only. This way no identity attributes will be disclosed. On the downside, this approach requires nifty algorithms to identify not only misspelled answers but also closely related answers to support convenient usability. Otherwise, re-purification of the real identity attributes might turn even for a legitimate requester very difficult or even impossible.

5.4 Conclusion

After looking at both applications of dilution as we implemented them (one of which employs active, the other one passive request analysis) from an adversaries perspective, we found that five different attack strategies where rather unsuccessful in attacking our passive polymorphic dilution implementation. While performing same tests against our active polymorphic implementation we identified some issues using real identity data for the purpose of dilution,

i.e., as virtual identity attributes. Nevertheless, both approaches appear to be promising alternatives to traditional privacy preserving technologies. Already our first proof-of-concept implementations turned out to be rather robust even from an attackers perspective.

Chapter 6

CONCLUSION

In this work we presented a novel concept attributable to the area of privacy preserving/enhancing technologies. The research this work is based on, was mainly driven by two goals:

1. Privacy in terms of confidentiality, by adding additional (irrelevant) information, i.e., dilution, instead of keeping information secret.
2. Decision on revealing personal information on a peer to peer base, depending on the knowledge/relationship between requester and the individual whose privacy is to protect.

Reaching out for these goals, raised a variety of challenges we had to tackle first. The most relevant research questions are listed in the following:

1. Which personal information is most significant in terms of online identifiability?
2. How does the online significance of personal information relate to demographic properties?
3. How robust is an online identity? Can it easily be tampered?
4. What information can be used for diluting personal information?
5. How to publish information, so the online identity can still be found?
6. What information can be used to distinguish/categorize different requesters?
7. How to compose a partial identity to be shown to the requester?

8. Is dilution applicable in real-world scenarios?
9. How robust is dilution (from an attackers perspective)?
10. What is the differentiator of dilution comparing to state-of-the-art?
11. How can individuals monitor their own online visibility without relying on services hosted by third parties?

6.1 Contribution

Answers to the before listed questions are provided throughout this work and provide the pillars for a new field in privacy related research. Particularly, the definitions and discussion on identity related data in Chapter 2 and the surveys on privacy related behavior (presented in Section 3.2), help us to understand and further discuss based on a common ground. The impact of online reputation was analyzed in Section 3.3.2 by our survey among human resource departments of four organizations/companies and in Section 3.3.3 by conducting a case study in which we position an imaginary identity within the Internet to confirm a curriculum vitae as we composed it. The results of these research efforts provided a solid base to build our own online reputation framework, which is presented in Section 3.3.4. To our knowledge this framework is the first approach to provide online reputation monitoring as an application instead of providing it as a service. This way any individual can run its own reputation monitoring and does not need to submit personal information to an online service which will then in turn monitor the online reputation. Among online services in this area, we already mentioned reputation defender [71], but there are several similar offerings, particularly in the business sector, such as BrandsEye [29], Brandtology [13], or Attentio [68].

However, the main contribution of our work is the design concept presented in Chapter 4, namely dilution. We present the concept in all details required for implementation and present two example applications, which we implemented as working proof-of-concepts. In one application we implemented the *passive request analysis* (as presented in Section 4.5.3) and in the other one we implemented the *active request analysis* (compare Section 4.5.3). This way we were able to not only have a full functional evaluation as outlined in Section 4.7, but furthermore also evaluate the robustness of our methodologies from an attackers perspective (see Chapter 5).

As we mentioned already in the introduction, we started from scratch with just as much inquiry on related work as it was necessary in order to ensure unique contribution comparing to state-of-the-art. Validation of the

uniqueness of our approach is given in Section 1.4, where we relate our contribution to other work, we found to be close enough to our methodology to be considered as relevant.

6.2 Future Work

While in this work and the preceding research we tried to look at a wide range of aspects in order to form the pillars for a new dimension in the area of privacy, we could only dig deeper on some of the aspects we considered. Here, we would appreciate to see further elaboration particularly on implementing and applying the concepts we suggested. Particularly the risk of harming reputation when dilution is applied and “bad identity attributes” are being used, poses a challenging research task, which probably needs more than computer science expertise.

Bibliography

- [1] Bug me not. <http://www.bugmenot.com>, last accessed 2012.
- [2] Markus Ackermann, Benjamin Ludwig, Krister Hymon, and Kai Wilhelm. Helloworld: An open source, distributed and secure social network. In *W3C Workshop on the Future of Social Networking*, 2009.
- [3] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Privacy Enhancing Technologies*, pages 36–58. Springer, 2006.
- [4] Anonymous S.A. Anonymouse.org. <http://anonymouse.org>, last accessed 2012.
- [5] appfield.net. Spambog. <http://www.spambog.com>, last accessed 2012.
- [6] ARD/ZDF. [ard-zdf-onlinestudie.de](http://www.ard-zdf-onlinestudie.de). <http://www.ard-zdf-onlinestudie.de/index.php?id=167>, last accessed 2012.
- [7] Marco Balduzzi, Christian Platzer, Thorsten Holz, Engin Kirda, Davide Balzarotti, and Christopher Kruegel. Abusing social networks for automated user profiling. In *Proceedings of the 13th international conference on Recent advances in intrusion detection*, RAID'10, pages 422–441, Berlin, Heidelberg, 2010. Springer-Verlag.
- [8] Christoph Bales. Datenspeicherung in Sozialen Netzwerken. Student's research project, University Mannheim, September 2009.
- [9] Anthony G. Basile. Tor-Ramdisk. <http://opensource.dyc.edu/tor-ramdisk>, last accessed 2012.
- [10] Mihir Bellare and Alexandra Boldyreva. The Security of Chaffing and Winnowing. In *Proceedings of the 6th International Conference on the Theory and Application of Cryptology and Information Security: Advances in Cryptology*, ASIACRYPT '00, pages 517–530, London, UK, UK, 2000. Springer-Verlag.

- [11] Leyla Bilge, Thorsten Strufe, Davide Balzarotti, and Engin Kirda. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *WWW*, pages 551–560. ACM, 2009.
- [12] Daniel Brandt. Scroogle. <http://www.scroogle.org>, last accessed 2012.
- [13] Brandtology. Brandtology. <http://www.brandtology.com>, last accessed 2012.
- [14] Bundesverband Deutscher Unternehmensberater. BDU-Personalberaterbefragung: Stellenwert von persönlichen Informationen im Internet für den beruflichen Erfolg nimmt weiter zu. Pressemitteilung. <http://www.presseportal.de/print.htx?nr=1073433>, October 2007.
- [15] Raymond B. Cattell. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.
- [16] M. Chew, D. Balfanz, and B. Laurie. (Under) mining Privacy in Social Networks. In *Proceedings of Web 2.0 Security and Privacy Workshop, W2SP*. Citeseer, 2008.
- [17] L. A. Cutillo, R. Molva, and T. Strufe. Safebook: A privacy-preserving online social network leveraging on real-life trust. *Comm. Mag.*, 47(12):94–101, December 2009.
- [18] Davide D’Amico and Dario Freni. FreeSBIE. <http://www.freesbie.org>, last accessed 2012.
- [19] Yvo Desmedt. Threshold Cryptography. In Henk C. A. van Tilborg and Sushil Jajodia, editors, *Encyclopedia of Cryptography and Security (2nd Ed.)*, pages 1288–1293. Springer, 2011.
- [20] Privoxy Developers. Privoxy. <http://www.privoxy.org>, last accessed 2012.
- [21] Dictionary.com LLC. Thesaurus. <http://www.thesaurus.com>, last accessed 2012.
- [22] dr.kaos, digunix, atlas, and beth. Anonym.OS. <http://sourceforge.net/projects/anonym-os/>, last accessed 2012.

- [23] Tung Dang Duc. Verwässerung des persönlichen Profils in Sozialen Netzwerken. Bachelor's thesis, University Mannheim, June 2010.
- [24] Manuel Egele, Andreas Moser, Christopher Kruegel, and Engin Kirda. PoX: Protecting users from malicious Facebook applications. *Comput. Commun.*, 35(12):1507–1515, July 2012.
- [25] Markus Engelberth, Felix Freiling, Jan Göbel, Christian Gorecki, Thorsten Holz, Ralf Hund, Philipp Trinius, and Carsten Willems. The InMAS Approach. *1st European Workshop on Internet Early Warning and Network Intelligence (EWNI)*, January 2010.
- [26] Markus Engelberth, Felix Freiling, Jan Göbel, Christian Gorecki, Thorsten Holz, Philipp Trinius, and Carsten Willems. Frühe Warnung durch Beobachtung und Verfolgung von bösartiger Software im Deutschen Internet: Das Internet-Malware-Analyse System (InMAS). *11. Deutscher IT-Sicherheitskongress*, May 2009.
- [27] Markus Engelberth, Felix C. Freiling, Jan Göbel, Christian Gorecki, Thorsten Holz, Ralf Hund, Philipp Trinius, and Carsten Willems. Das Internet-Malware-Analyse-System (InMAS) - Ein System zur großflächigen Sammlung und Analyse von Schadsoftware im Internet. *Datenschutz und Datensicherheit*, 35(4):247–252, April 2011.
- [28] Markus Engelberth, Jan Gobel, Christian Gorecki, and Philipp Trinius. Mail-Shake. In *Proceedings of the 2009 20th International Workshop on Database and Expert Systems Application*, DEXA '09, pages 43–47, Washington, DC, USA, September 2009. IEEE Computer Society.
- [29] Exclusive Access Trading 845 (Pty) Ltd. BrandsEye. <http://www.brandseye.com>, last accessed 2012.
- [30] Facebook, Inc. Facebook. <http://www.facebook.com>, last accessed 2012.
- [31] face.com. face.com. <http://face.com>, last accessed 2012.
- [32] Adrienne Felt and David Evans. Privacy Protection for Social Networking Platforms. In *Proceedings of W2SP 2008: Web 2.0 Security and Privacy*, May 2008.
- [33] J. Fogel and E. Nehmad. Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in Human Behavior*, 25(1):153–160, January 2009.

- [34] Agriculture and Consumer Protection German Federal Ministry of Food. Umfrage zu Haltung und Ausmaß der Internetnutzung von Unternehmen zur Vorauswahl bei Personalentscheidungen. <http://www.bmelv.de/SharedDocs/Downloads/Verbraucherschutz/InternetnutzungVorauswahlPersonalentscheidungen.htm>, July 2009.
- [35] Pascal Göbel. Clusterbasierte Analyse zur Internetpräsenz von Personen. Master's thesis, University Mannheim, June 2010.
- [36] Google, Inc. Google. <http://www.google.com>, last accessed 2012.
- [37] Google, Inc. Picasa. <http://picasa.google.com>, last accessed 2012.
- [38] Christian Gorecki, Felix C. Freiling, Marc Kühner, and Thorsten Holz. TRUMANBOX: improving dynamic malware analysis by emulating the internet. In *Proceedings of the 13th international conference on Stabilization, safety, and security of distributed systems*, SSS'11, pages 208–222. Springer-Verlag, October 2011.
- [39] Martin Gräßlin. MailShake Mailclient Plugin. Master's thesis, University Mannheim, 2010.
- [40] Johannes Grohmüller. Angriffe auf dynamische Profilgeneratoren. Bachelor's thesis, University Mannheim, February 2011.
- [41] Louis Guttman. Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2):149–161, June 1954.
- [42] Marit Hansen, Markus Hansen, Marita Häuser, Kai Janneck, Henry Krasemann, Martin Meints, Sebastian Meissner, Maren Raguse, Martin Rost, and Jan Schallaböck. Verkettung digitaler Identitäten: Untersuchung im Auftrag des Bundesministeriums für Bildung und Forschung. *Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein*, October 2007.
- [43] John A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 99th edition, 1975.
- [44] Matt Hawkins. Antiphorm. <http://www.antiphorm.co.uk>, last accessed 2012.
- [45] Thorsten Holz, Christian Gorecki, Konrad Rieck, and Felix C. Freiling. Measuring and Detecting Fast-Flux Service Networks. In *NDSS*. The Internet Society, 2008.

- [46] Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.
- [47] Nick Jones. PHP-Fusion. <http://www.php-fusion.co.uk>, last accessed 2012.
- [48] Alexander Juhn. Reputation Monitoring - Entwicklung eines halbautomatischen Systems zu Überwachung der eigenen Internetpräsenz. Master's thesis, University Mannheim, December 2011.
- [49] Henry Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.
- [50] Christoph Klasik. Alternative Ansätze zum Schutz der Privatsphäre. Student's research project, University Mannheim, January 2011.
- [51] Daniel Köhler. Fiktive Identitäten in Sozialen Netzwerken: Chancen und Risiken im Bereich Recruiting. Student's research project, University Mannheim, August 2010.
- [52] Sinem Kuz. Verwässerung von Persönlichkeitsprofilen im Internet. Master's thesis, RWTH Aachen, September 2009.
- [53] Phoenix Labs. PeerGuardian 2. <http://sourceforge.net/projects/peerguardian/>, last accessed 2012.
- [54] Scott R. Lemmon. Proxomitron. <http://www.proxomitron.info>, last accessed 2012.
- [55] LinkedIn Corporation. LinkedIn. <http://www.linkedin.com>, last accessed 2012.
- [56] J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
- [57] P.K. Manoharan. *Education And Personality Development*. APH Publishing Corporation, 2008.
- [58] Boris Margara. Identifizierbarkeit im Internet: Zur Signifikanz personenbezogener Daten. Bachelor's thesis, University Mannheim, March 2009.
- [59] Microsoft Corporation. Bing.com. <http://www.bing.com>, last accessed 2012.

- [60] F. Moos. *Datenschutzrecht-schnell erfasst*. Springer, 2006.
- [61] Mozilla Foundation. Firefox. <http://www.mozilla.org/en-US/firefox/new/>, last accessed 2012.
- [62] MyHeritage (UK) Ltd. Verwandt.de. <http://www.verwandt.de>, last accessed 2012.
- [63] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *2009 30th IEEE Symposium on Security and Privacy*, pages 173–187. IEEE, 2009.
- [64] Konrad Nuhn. Einblicke von innerhalb und außerhalb in Soziale Netzwerke. Student’s research project, University Mannheim, May 2010.
- [65] David Passarelli. Analyse des Vorgehens und Verhaltens von Spammern und Harvestern bei dem Auffinden von E-Mail Adressen im Internet. Student’s research project, University Mannheim, February 2010.
- [66] Alexander Pfister. Modellierung und Analyse von Fraud in elektronischen Geschäftsprozessen. Master’s thesis, University Mannheim, April 2009.
- [67] Andreas Pfitzmann and Marit Hansen. Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management - A Consolidated Proposal for Terminology. http://dud.inf.tu-dresden.de/Anon_Terminology.shtml, February 2008.
- [68] Progressive Media Group PLC. Attentio. <http://attentio.com>, last accessed 2012.
- [69] Freenet Project. Freenet - The Free Network. <http://freenetproject.org>, last accessed 2012.
- [70] Katharina Reich. Online Reputation Inspection: Continuous Monitoring the Online Presence of People. Bachelor’s thesis, University Mannheim, June 2011.
- [71] Reputation.com, Inc. Reputation Defender. <http://www.reputation.com>, last accessed 2012.
- [72] Michael Riecker. Forensische Datenanalysen mit Data Mining. Master’s thesis, University Mannheim, March 2009.

- [73] Ronald L. Rivest. Chaffing and Winnowing: Confidentiality without Encryption. <http://theory.lcs.mit.edu/~rivest/chaffing.txt>, March 1998.
- [74] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, 1998.
- [75] Patrick Scharrenberg. Analyzing Fast-Flux Service Networks. Master's thesis, RWTH - University of Aachen, 2008.
- [76] Jörg Schieb and Mirko Müller. *Meine Daten schützen*, volume 1. Stiftung Warentest, 2008.
- [77] Matthias Schwenke. *Individualisierung und Datenschutz*. Deutscher Universitäts-Verlag, June 2006.
- [78] U.S. Bureau of Industry Security. Electronic Code of Federal Regulations. <http://www.ecfr.gov>, last accessed 2012.
- [79] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: discovery and applications of usage patterns from Web data. *SIGKDD Explor. Newsl.*, 1(2):12–23, January 2000.
- [80] Nina Sophie Stadler. E-Recruitment. Student's research project, University Mannheim, September 2010.
- [81] Surfboard Holding BV. Ixquick. last accessed 2012.
- [82] Latanya Sweeney. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, July 2002.
- [83] The Tor Project, Inc. Tor: Anonymity Online. <http://www.torproject.org>, last accessed 2012.
- [84] Vincent Toubiana, Lakshminarayanan Subramanian, and Helen Nissenbaum. TrackMeNot: Enhancing the privacy of Web Search. *CoRR*, abs/1109.4677, 2011.
- [85] TU Dresden. Cookie Cooker. <http://www.cookiecooker.de>, last accessed 2012.

- [86] W3C. Platform for Privacy Preferences (P3P) Project, 2006. <http://www.w3.org/P3P>.
- [87] Christoph Wachter and Mathias Jud. Picidae. net.picidae.net, last accessed 2012.
- [88] Gilbert Wondracek, Thorsten Holz, Engin Kirda, and Christopher Kruegel. A Practical Attack to De-anonymize Social Network Users. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, SP '10, pages 223–238. IEEE Computer Society, May 2010.
- [89] Ying-Ming Wu, Hsueh-Wu Wang, Yen-Ling Lu, Shin Yen, and Ying-Tung Hsiao. Facial Feature Extraction and Applications: A Review. In Jeng-Shyang Pan, Shyi-Ming Chen, and Ngoc Thanh Nguyen, editors, *ACIIDS (1)*, volume 7196 of *Lecture Notes in Computer Science*, pages 228–238. Springer, 2012.
- [90] Xing AG. Xing. <http://www.xing.com>, last accessed 2012.
- [91] Yahoo! Inc. Altavista.com. <http://www.altavista.com>, last accessed 2012.
- [92] Yahoo! Inc. Yahoo! <http://www.yahoo.com>, last accessed 2012.
- [93] yasni GmbH. Yasni. <http://www.yasni.de>, last accessed 2012.

Appendix A

Related Solutions, Products, and Services

Complete list of the offers we evaluated:

- Services
 - Anonymous.org [4]
 - Bugmenot.com [1]
 - Spambog.com [5]
 - Ixquick.com [81]
 - Picide.net [87]
 - Scroogle.org [12]
- Programs
 - Cookiecutter [85]
 - Proxomitron [54]
 - Privoxy [20]
 - Tor [83]
 - Freenet (Darknet) [69]
 - PeerGuardian [53]
- Operating Systems
 - Anonym.OS [22]
 - FreeSBIE [18]
 - Tor Ram Disk [9]

Appendix B

Survey

Questions as used in the survey.

#	Question	Possible Answers
1	Age	< 14 $14 - 17$ $18 - 21$ $22 - 29$ $30 - 49$ > 49
2	Gender	female male
3	Job description	IT-related Media, Communication & Journalism Art, Culture & Fashion Tourism, Activity & Sport Mathematics, Physics, Biology & Chemistry Health-care & Medical Science Geography & Geology Environmental/Social Science & Education Engineering, Automotive & Metal Clerk & Temporary Employee Office, Financial, Economy & Law Other: -----
	Online time per day	
<i>continue next page</i>		

<i>continuation of previous page</i>		
#	Question	Possible Answers
4	Job-related	$< 1h$ $1 - 2h$ $2 - 3h$ $3 - 4h$ $> 4h$
5	Private	$< 1h$ $1 - 2h$ $2 - 3h$ $3 - 4h$ $> 4h$
	How do you estimate your sensitivity in sharing the following information (1: low, 5:high)	
6	Name	1 – 5
7	City	1 – 5
8	Street	1 – 5
9	Email (business)	1 – 5
10	Email (private)	1 – 5
11	Messenger Id	1 – 5
12	Job description	1 – 5
13	Hobbies	1 – 5
14	Phone-No.	1 – 5
15	Pictures (of yourself)	1 – 5
	How do you estimate our online visibility (1: high, 5: low)	
16	Job-related	1 – 5
17	Private	1 – 5
	How important do you estimate online reputation (1: high, 5: low)	
18	Job-related	1 – 5
19	Private	1 – 5
	How important do you estimate YOUR online reputation (1: high, 5: low)	
<i>continue next page</i>		

<i>continuation of previous page</i>		
#	Question	Possible Answers
20	Job-related	1 – 5
21	Private	1 – 5
	Do a Google search using your last name as a search term. Look at the first 10 results.	
22	How many refer to you	1 – 10
23	How many of these hits refer to you privately	1 – 10
24	How many of these hits refer to you job-related	1 – 10
	Do a Google search using your first and last name as a search term. Look at the first 10 results.	
25	How many refer to you	1 – 10
26	How many of these hits refer to you privately	1 – 10
27	How many of these hits refer to you job-related	1 – 10
	Do a Google search using your first and last name and your city as a search term. Look at the first 10 results.	
28	How many refer to you	1 – 10
29	How many of these hits refer to you privately	1 – 10
30	How many of these hits refer to you job-related	1 – 10
	Do a Google search using your first and last name and your company/university/etc. as a search term. Look at the first 10 results.	
31	How many refer to you	1 – 10
<i>continue next page</i>		

<i>continuation of previous page</i>		
#	Question	Possible Answers
32	How many of these hits refer to you privately	1 – 10
33	How many of these hits refer to you job-related	1 – 10
	Do a Google search using your job-related email as a search term. Look at the first 10 results.	
34	How many refer to you	1 – 10
35	How many of these hits refer to you privately	1 – 10
36	How many of these hits refer to you job-related	1 – 10
	Do a Google search using your most used private email address as a search term. Look at the first 10 results.	
37	How many refer to you	1 – 10
38	How many of these hits refer to you privately	1 – 10
39	How many of these hits refer to you job-related	1 – 10
	Do a Google search using your most used user-name (e.g. online-shopping, instant-messenger, etc.) as a search term. Look at the first 10 results.	
40	How many refer to you	1 – 10
41	How many of these hits refer to you privately	1 – 10
42	How many of these hits refer to you job-related	1 – 10
<i>continue next page</i>		

<i>continuation of previous page</i>		
#	Question	Possible Answers
43	How many of these link to further personal information (e.g. address) about you	1 – 10
44	Does the result of the before queries surprise you?	Not surprising Surprising Very surprising
	Visit http://www.verwandt.de/karten/ and enter your last name into the search form.	
45	How many persons in Germany share the same name with you?	0 – 250 251 – 1000 1001 – 5000 5001 – 25000 25001 – 100000 100001 – 250000 > 250000
46	Is your first name listed among the top ten?	Yes No
	Visit http://www.yasni.de and enter your first and last name	
47	How many of the returned results refer to you	0 1 – 5 6 – 20 > 20
48	Do you use Internet platforms (e.g. online social networks) to publish information about yourself	Yes, private only Yes, job-related only Yes, private and job-related No, not at all
<i>continue next page</i>		

<i>continuation of previous page</i>		
#	Question	Possible Answers
49	Please choose all online social networks where you have an account	Studi-/Mein-/SchülerVZ Facebook Wer-kennt-wen Myspace Lokalisten Stayfriends ICQ Yasni XING Others
50	Did you customize your privacy settings	Yes No
51	How is your attitude regarding privacy within online social networks (1: I don't care, 5: Very important)	1 – 5
52	How truthful are your details in online social networks (1: no truthful details, 5: only truthful details)	1 – 5
53	How detailed is your provided information in online social networks (1: very detailed, 5: minimal details only)	1 – 5
	Do a Google Image query using your first and last name as a search term. Look at the first 18 results.	
54	How many of the first 18 returned pictures show you	0 1 – 3 4 – 7 > 8
55	Do the pictures show you in rather private or job-related context	Private Job-related
<i>continue next page</i>		

<i>continuation of previous page</i>		
#	Question	Possible Answers
56	How many of the first 18 pictures may harm/hamper your carrier	0 1 – 3 4 – 7 > 8
57	Estimate how many pictures showing you can be found via social networks	0 1 – 3 4 – 7 8 – 20 > 20
	How many of these pictures show you in the following situations (or similar)	
58	Party	0 1 – 3 4 – 7 8 – 20 > 20
59	Holidays	0 1 – 3 4 – 7 8 – 20 > 20
60	Free time	0 1 – 3 4 – 7 8 – 20 > 20
61	Sport	0 1 – 3 4 – 7 8 – 20 > 20
62	Job	0 1 – 3 4 – 7 8 – 20 > 20
<i>continue next page</i>		

<i>continuation of previous page</i>		
#	Question	Possible Answers
	Any comments can be submitted in the following field (not part of the survey)	<div>-----</div> <div>-----</div> <div>-----</div> <div>-----</div>

Appendix C

Interview Questions

The following questions were asked during the interviews with recruiters of four different companies/organizations. Note that the interviews were in German, so the questions below are translations of the original questions.

1. Which channels do you use for recruiting in your organization?
 - Announcement in newspaper
 - Announcement on homepage
 - Online application form
 - Open application
 - Search via human resource agencies
 - Search in job portals
 - Search CV database
2. What are the advantages of your preferred channel?
3. Describe the recruiting process for the channels you exploit.
4. Do you have particular requirements a candidate has to meet?
5. How do you assess the skills of your candidates?
6. Do you use the Internet for a background check on your candidates?
 - (a) If yes
 - Where are you looking for information?
 - Do you gather additional information in advance or afterwards?

- Do you validate statements in the CV? How do you proceed here?
 - What do you expect from this information?
 - What are the consequences, in case you find additional information online?
- (b) If no
- Why not?